基于信息资源聚合的交互式报告生成模型研究

袁雪[1,2] 刘敏娟[1,2] 刘洪冰[1,2] 王新[1,2] 赵婉婧[1,2] 江浩[1,2]
(1. 中国农业科学院农业信息研究所，北京 100081；2. 农业农村部农业大数据重点实验室，北京 100081)

摘要：首次提出基于信息资源聚合的交互式报告生成模型，该模型的应用旨在探索一种新的信息聚合服务方式，促进用户对特定领域信息资源的快速获取和高效利用。本文首先从面向用户需求的信息资源知识体系构建、信息资源遴选采集、信息资源整理序化、信息资源细粒度聚合4个方面构建模型，然后对报告生成的关键环节进行分步骤解析，最后以生成茶产业对外合作发展报告为例进行原型实现，详述报告生成过程及主要功能，实证表明能够明显减少领域用户获取多来源渠道信息的时间，提升信息输入的效率；能够一键生成定制化报告，操作简单便捷；报告中段落可按国家、年份等多维度检索查询，提升查找资料或素材的效率。

关键词：信息聚合；信息单元；信息组织；细粒度；交互式报告
中图分类号：G250　　DOI：10.3772/j.issn.1673-2286.2020.10.006
引文格式：袁雪，刘敏娟，刘洪冰，等. 基于信息资源聚合的交互式报告生成模型研究[J]. 数字图书馆论坛，2020 (10)：38-43.

现今海量信息资源为用户带来丰富数据和信息的同时，也带来了信息超载的困难，不利于信息的高效获取与利用[1]。随着信息组织技术的发展，用户更多地希望帮助他们实现信息资源跨语种、跨载体、跨领域的有机关联，按照特定需求从大量分散来源中获取信息并以序化整合的方式提供一站式信息服务。为实现用户快速、高效、多方位获取有序的、成体系信息的意愿，交互式报告作为一种新型的信息服务方式应运而生，弥补了传统编辑报告在时效性、便捷性、个性化等方面的不足。其实施过程中的PDF文档碎片化，是指识别PDF文档中章节、图表、段落等细颗粒信息单元，提取文本阅读顺序并对文档结构进行层次分析，将其分解成一个有层次、有逻辑的有机体，是实现细粒度信息单元重组和深度知识挖掘的基础；继而利用非传统的编辑、加工与生成方式，融合PDF文档碎片化技术与交互式操作，将相关领域横纵向的专业信息资源，按照相应的知识资源体系进行规模化地获取、遴选与汇聚，经细粒度加工、深层次揭示，实现动态重组与发布。实施效果表明：能显著提升用户信息输入的效率，实现有针对性的、轻量化的阅读；能够一键生成定制化报告，支持多维度分面检索，提高查找信息的效率。

1 交互式报告生成模型理论基础

1.1 聚合单元划分

Sandusky等[2]认为文献资源包括两种类型的结构，一种是如摘要、正文、图表和参考文献等的形式结构；另一种是文献组织成一个叙述部分，如文献综述、研究方法、结果与讨论等反映文献构思的逻辑结构。据此，本文以形式结构和逻辑结构的划分作为聚合单元划分的依据。形式结构分析中将篇章视为整体，其各个部分视为组成要素。参考顾小清[3]提出的通过分析期刊论文的形式结构而拆分期刊论文的组成部分，文中将PDF文档的外部特征如标题、著者、关键词等进行结构化存储，其中针对形式结构缺乏严谨和统一的资源进

行处理后结构化著录。逻辑结构分析针对的是篇章形式结构中的正文部分，文献的逻辑结构包括两个方面：一是作者根据行文框架与逻辑对整篇文档进行分割，即节段单元；二是具有一定交际意图和修辞目的的语篇结构，即句群单元[4]。考虑到语义功能的标注完整性及以用户较为希望的聚合单元粒度，采用以正文部分篇章节段单元和图表单元作为分割的颗粒度，以方便用户快搜检索定位所需的单元内容，从而节省浏览和查找其他不相关信息的时间。

1.2 信息资源聚合

信息资源聚合被认为是网络环境下知识组织的一种新模式，正成为信息组织与检索领域的研究热点[5]，其主要是对数字或网络环境下文本信息单元中包含的相关信息进行抽取与重组。通过对国内外研究梳理发现，近年来的研究主要围绕信息聚合相关理论体系、聚合技术方法、聚合效果显现评估、聚合应用实践等方面[1, 6-8]，并取得了一定的研究成果。细粒度信息资源聚合作为信息资源聚合的主要模式之一[9]，相关研究聚焦聚合单元元数据、多维度语义标注机制、多维语义聚合等，为细粒度聚合单元的解析、抽取、分析与管理提供相应的理论依据。本文在前人研究基础上进行拓展与延伸，并着重结合用户信息需求，从优化信息资源组织角度进行信息的整合序化，为用户高效便捷地获取与利用信息提供服务：其中元数据框架的制定参考曹树金等[1]构建的文献标识符、关键词、来源等核心元素，以及标题、责任者、日期、语种资源等篇章方面的描述元素作为资源元数据标准；聚合单元的多维度语义标注参考胡潜等[9]提出的从行业主题和用户群体两个维度进行组织体系的构建与标注，依据面向用户需求的资源知识体系进行基于机器学习的文本自动分类标注，实现基于机器学习的段落文本特征支持向量机模型和随机森林模型的信息细粒度加工和动态重组。

2 交互式报告生成模型构建

交互式报告生成模型的构建主要包括面向用户需求的信息资源知识体系构建、信息资源遴选采集、信息资源整理序化、信息资源细粒度聚合4个构成要素，具体如图1所示。该模型依据构建的资源知识体系，将文档中段落及图表等细颗粒度的信息单元进行解构、重组与主题聚合，最大限度地保留了段落上下文的行文逻辑，报告内容丰富多元，易用且易读。报告的动态性体现在完整保留历史信息的同时完成最新资源的及时、动态更新以及多维度检索查询等多种用户交互功能，实现了专业信息资源的遴选汇聚、细粒度加工、深层次揭示、信息重组以及一键式报告生成等。

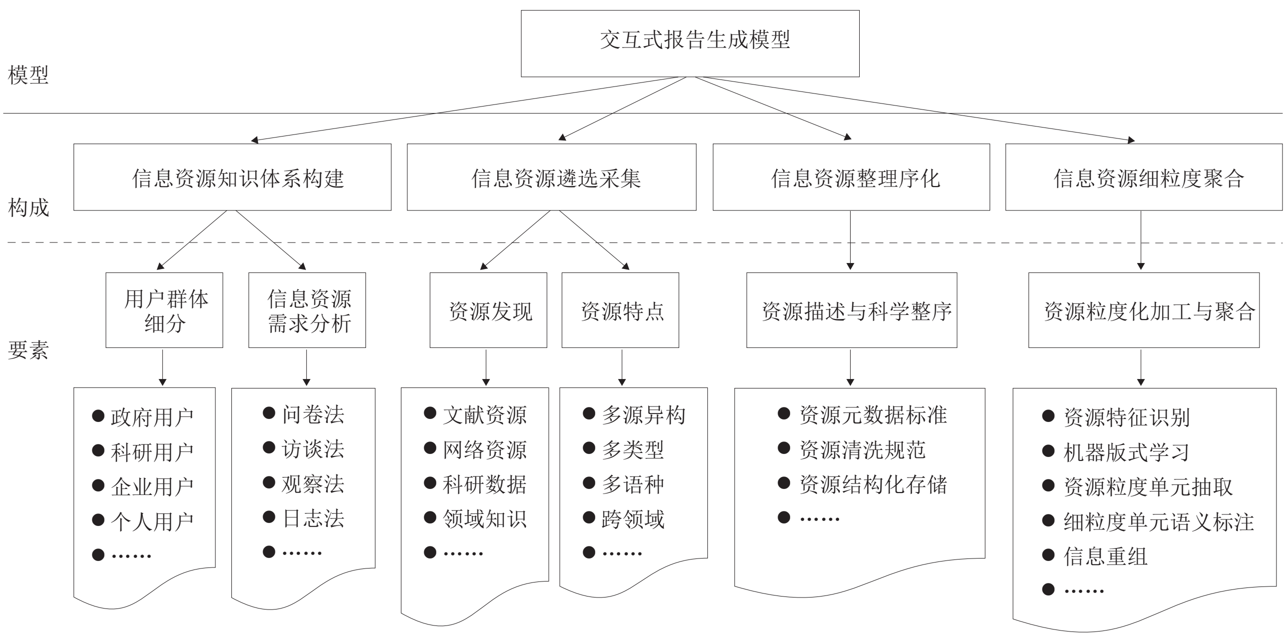

图1　交互式报告生成模型构成与要素

2.1 信息资源知识体系构建

信息资源知识体系构建以用户需求为出发点和落脚点，形成一个完整树状组织结构，用于报告中待聚合资源遴选的依据和聚合细粒度信息单元的语义标注。其中用户信息需求源于其职业、角色、工作内容等，直接决定所需信息的范围、类型、数量和深度。为了满足多类型用户通用信息需求及有代表性的个性化需求，需要对用户群体进行细分。细分用户群体需求可能存在差异，具体来说，用户组成涉及各行各业，覆盖面较广，按工作性质分为政府用户、科研用户、企业用户和个人用户等，各类用户需求侧重点不同。如政府用户普遍关注宏观政策、政府管理等方面的信息；科研用户往往关注学术与研究方面的信息，且兼具广度与深度；企业用户则更多关注行业发展、市场动态、商业资讯等；个人用户关注内容较为多样且分散，因此考虑作为个体进行补充，不作为用户细分群体的重点。上述用户群体的信息需求采用问卷法、调查法、观察法、日志法进行收集，同时对上述内容进行整理与分析，作为资源知识体系构建的依据。

从前期调研情况来看，各类型用户的信息需求主要涉及相关主题领域信息获取的高效性以及信息利用的便捷性。信息获取的高效性在于信息的规模化获取、遴选与汇聚，以及有效的、分门别类的内容组织；信息利用的便捷性在于有针对性的、精准的、无障碍的轻量化阅读。据此，依据用户需求和应用场景，收集、分析用户需求信息，按照信息资源分面分类的方法，横向延伸，纵向深入，形成面向用户需求的资源知识体系。纵向等级是树形结构，横向展开是平行类目，可依据用户需求变化进行动态调整，各层级父主题下设若干个子类，直至满足用户信息需求为止。信息资源知识体系可用于报告内容发布模板设置，即生成动态报告导航目录。除此之外，应按照用户需求设置一键生成定制化报告，支持细颗粒信息单元可按国家、年份等进行多维度检索查询，以方便用户信息查找以及提高阅读效率。

2.2 信息资源遴选采集

信息资源遴选采集是按照构建的资源知识体系主题特征对无序聚合候选资源进行筛选过滤。资源遴选对象包括文献资源、网络资源、科研数据和领域知识等，遴选过程是从多源异构、多类型、多语种、跨领域的资源中优选具有权威性、客观性、代表性的资源作为待聚合的资源，并利用相关信息采集技术实现快速有效的信息获取。具体来说，采集内容围绕构建好的资源知识体系的主题领域遴选可靠的信息资源，采集其元数据及PDF全文，非PDF文档进行转换处理，并整合存储到创建好的数据库中。采集渠道均来自互联网，考虑到上述资源对用户的开放程度，一般采用公开获取的权威资源作为首选，主要来自国际组织官网、开放获取数据库、公开的报告与文献等，一方面保证了资源的可用性；另一方面由于这些资源经过专业人员严格的遴选与评估，也保证了资源的完整性、可靠性以及权威性。具体包括：①网站类，如国际组织网站、政府机构网站、高校网站、行业协会网站、行业垂直网站等；②数据库类，如开放获取数据库、商业数据库等提供的知识资源；③报告与文献类，如国际统计机构的调查资料、市场调研报告、新闻报道、学术期刊等。标准化的模板包括标题、作者、发布年、关键词等元数据项及全文文档的信息录入。

2.3 信息资源整理序化与细粒度聚合

信息资源整理序化是将采集过程中大量的信息资源进行清洗规范，实施统一的资源描述并进行结构化存储，变无序资源为有序资源。该过程中涉及资源元数据标准的建立、资源清洗规范操作流程与方式的建立、资源存储数据库的建立等相关工作，从而实现资源的科学整序。信息资源细粒度聚合由资源粒度化抽取、细粒度单元语义标注和细粒度单元信息重组三部分构成。资源粒度化抽取实质上是聚合信息资源解构的过程，鉴于文本资源主要以较常见的PDF格式存储为主[10-11]，其他WORD、TXT或网络信息片段也可转换为PDF格式，因此粒度化抽取对象选取PDF版式文档，方法采用基于机器学习的段落文本特征支持向量机模型和随机森林模型[12]，根据文章行文框架与逻辑结构对整篇文档进行分割，预测目标PDF全文的标题、章节和图表等结构信息，抽取以段落或图表为最小颗粒度的信息单元，信息量大且多元、广泛而丰富，细粒度信息单元实时动态更新，历史信息与最新信息同步留存，且段落中上下文行文逻辑完整保留；PDF解构后的段落与图表等更细颗粒度的信息单元，以XML格式在数据库中进行结构化存储，操作对象包括文本内容的结构、属性与关联信息等。细粒度语义标注采用基于机器学习的文本

自动分类方法按照面向用户需求的资源知识体系进行分类标注[13]，具体是先将标注好的PDF文档碎片化信息单元作为训练集，通过机器学习算法从文本中整理出能够有效分类的规则，生成分类器，将生成的分类器应用在有待分类的文本集合中，实现自动分类标注，同时为了增强自动分类结果的精度，添加人工校改的辅助功能，便于对分类结果进行校准。细粒度单元信息重组是碎片化后的细粒度信息单元聚合重构的过程，实时将带有语义的单元信息按照一定的资源知识体系进行重新组合成拥有新内容构成的报告，按此循环往复而完成信息的全面汇聚。

3 交互式报告服务实践

服务实践以生成茶产业对外合作发展报告为例进行原型系统实现，解析报告生成过程及主要功能，关键实施步骤如图2所示。文中茶产业指中华人民共和国农业行业标准《农产品分类与代码》（NY/T3177—2018）种植业产品的饮料作物产品中的茶叶进行报告原型的服务实现。原型功能设计的服务内容包括：①面向用户的一键式动态报告生成，即用户按照报告的目录导航，选择感兴趣的内容，通过一键式操作立即生成定制化报告，提升获取报告的便捷性；②基于聚合细粒度信息资源的多维检索查询服务，即报告中段落信息依据国家与年份等进行分面检索，提升资源的专指性，并支持自定义导出到本地文件。

茶产业对外合作发展报告综合政府用户、企业用户、科研用户、个人用户4类不同用户的专业背景和需求内容，构建多层次的报告目录导航，构建过程以能够向用户提供所需知识为准则，以知识的粒度能恰当解决当前问题或需求为目标，使报告层次结构能够满足用户全方位、多层次的知识需求。依据用户需求分析的综合结果，构建了包含一级主题6个（发展概述、供需形势、进出口分析等）、二级主题22个（产业链分析、市场需求、市场价格等）的导航目录，如图3所示。可按照用户的实际需求进行一键式定制化报告生成，可支持全选整个报告目录的内容，以及部分选择一级主题或二级主题的内容。

待聚合信息资源遴选采集的范围按网站类、数据库类、报告与文献类分别进行优选，资源渠道包括网站类的资源主要取自美国农业部（USDA）、联合国粮农组织（FAO）、中国农业农村部、世界茶进出口委员

图2　茶产业对外合作发展报告生成关键环节解析

图3　茶产业对外合作发展报告导航目录

会等；数据库类的资源主要取自文献数据库、Science Direct、Stasita、美国经济统计数据库等；报告与文献类资源取自国际统计机构调查报告、市场贸易期刊、行业咨询报告等公开资料。以上信息资源通过半自动化采集方式获取PDF全文与元数据，非PDF文档可以统一转

换为PDF进行处理，并经过科学整理与序化。

满足多方位用户需求不仅需要广泛地采集资源，还要对已有的资源进行开发性组织，使资源再生为新的信息或知识。为了更好地实现待聚合信息资源的再利用，需经前述的信息资源粒度化加工和信息资源细粒度聚合，从而实现茶产业对外合作发展报告生成。报告可于多终端上线发布，支持一键定制化报告（见图4），细粒度段落的国家与年份多维度分面检索（见图5）。

图4 茶产业对外合作发展报告一键式生成

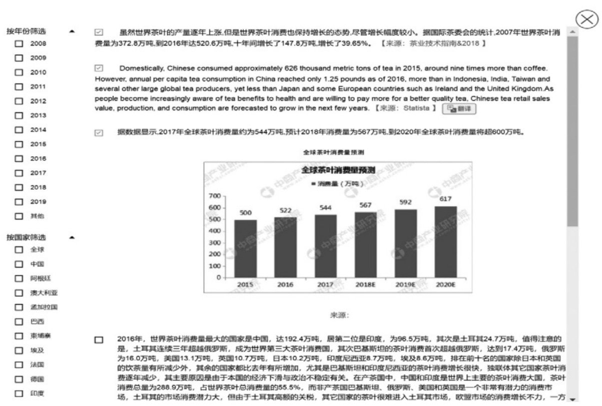

图5 茶产业对外合作发展报告按国家、年份等多维度分面检索

4 结语

学者研究指出用户在一般情况下查找所需信息并非要获得整篇文档，更多的是相互关联的信息片段或知识元，而信息片段的分散性导致很难较为全面地获取、描述与揭示，因此用户常需针对相关主题遍历查找、浏览、提取、整合等多项操作才能实现信息的有效输入，这一过程耗时费力，且较易在信息搜寻过程中产生“信息迷航”现象[14]。现阶段为提升用户获取信息的效率，信息聚合研究正成为图书情报领域关注的热点，然而学界对其研究至今仍然较多集中在聚合理论框架、机制模型、方法技术方面，应用层面的研究相对较少且有一定局限性[1]。交互式报告生成模型的设计提出是信息聚合领域研究成果解决社会实际问题的应用实践，该方法能够较好地帮助用户梳理多来源相关主题信息资源，并支持一键式定制化报告、多维度分面检索、英中互译等，大幅度缩减了用户获取多源异构信息资源的时间，同时优化了用户交互式体验效果。文中以茶产业对外合作发展报告为例，展示了原型系统的实施效果，体现出较好的适用性，该方法仍适用于其他主题领域报告的生产与实现，将有效促进且提升用户信息资源获取组织的效率与效用。

参考文献

[1] 曹树金，李洁娜，王志红. 面向网络信息资源聚合搜索的细粒度聚合单元元数据研究 [J]. 中国图书馆学报，2017，43（4）：74-92.

[2] SANDUSKY R J，TENOPIR C. Finding and using journal-article components：impacts of disaggregation on teaching and research practice [J]. Journal of the American Society for Information Science and Technology，2008，59（6）：970-982.

[3] 顾小清. 终身学习视野下的微型移动学习资源建设 [M]. 上海：华东师范大学出版社，2011：78-80.

[4] 曹树金，马翠嫦. 信息聚合概念的构成与聚合模式研究 [J]. 中国图书馆学报，2016，42（3）：4-19.

[5] 邱均平，王菲菲. 基于共现与耦合的馆藏文献资源深度聚合研究探析 [J]. 中国图书馆学报，2013（3）：25-33.

[6] 毕强，尹长余，滕广青，等. 数字资源聚合的理论基础及其方法体系建构 [J]. 情报科学，2015，33（1）：2，9-14.

[7] 王晓娟，谭建龙，刘燕兵，等. 基于自动机理论的PDF文本内容抽取 [J]. 计算机应用，2012，32（9）：2491-2495.

[8] BESSAI-MECHMACHE F Z，ALIMAZIGHI Z. Aggregated search in XML documents [J]. Journal of Emerging Technologies in Web Intelligence，2012，4（2）：181.

[9] 胡潜，李静. 面向用户的行业信息资源聚合研究——以母婴健康行业用户知识社区为例 [J]. 图书情报知识，2018（1）：87-94.

[10] ZHANG L. Grasping the structure of journal articles：Utilizing the functions of information units [J]. Journal of the American Society for Information Science and Technology，2012，63（3）：469-480.

[11] 王洪宪，王玉玫. 网络信息资源与大学生利用研究 [M]. 北京：中国财政经济出版社，2014：247-254.

[12] 段飞虎，吴盼盼，冯自强，等. 一种基于机器学习的论文碎片化信息抽取方法［P］. CN108536683A，2018-09-14.
[13] 邸锦. 基于支持向量机的文本分类问题的研究［D］. 北京：北京交通大学，2008.
[14] 叶鹏. 基于机器学习的中文期刊论文自动分类研究［D］. 南京：南京大学，2013.

作者简介

袁雪，女，1984年生，博士，馆员，研究方向：信息资源组织与管理。
刘敏娟，女，1982年生，博士，副研究馆员，通信作者，研究方向：信息资源组织与管理，E-mail：liuminjuan@caas.cn。
刘洪冰，女，1990年生，硕士，馆员，研究方向：用户需求分析。
王新，男，1986年生，博士，馆员，研究方向：信息组织技术与实践。
赵婉婧，女，1987年生，硕士，助理研究员，研究方向：信息组织技术与实践。
江浩，男，1974年生，硕士，工程师，研究方向：信息资源分析。

Model Research of Interactive Reports on Fine-grained Aggregation of Information Resource

YUAN Xue[1,2] LIU MinJuan[1,2] LIU HongBing[1,2] WANG Xin[1,2] ZHAO WanJing[1,2] JIANG Hao[1,2]
(1. Agricultural Information Institute, Chinese Academy of Agricultural Science, Beijing 100081, China; 2. Key Laboratory of Big Agri-data, Ministry of Agriculture and Rural Areas, Beijing 100081, China)

Abstract: Method of generating interactive reports in response to user needs are proposed for the first time, and generation methods for dynamic and interactive reports aim to explore novel approaches of services in information aggregation, thereby facilitating rapid acquisition and efficient utilization of information resources in a certain field for users. In this paper, a model to generate dynamic and interactive reports was firstly constructed based on five elements oriented at users' demands, i.e., design of resources knowledge, selection and collection of information resources, information resources sorting and ordering, fine-grained aggregation and service applications. Afterwards, key processes in report generation system were analyzed step by step. Finally, a prototype system was realized by taking the generation of a development report on foreign cooperation in the tea industry as an example, during which the generation process and main functions were discussed in detail. The demonstration shows that the time to obtain information from multiple channels for domain users is significantly reduced, thereby enhancing the efficiency of information input. A customized report can be generated with one click, signifying simple and convenient operation. Also, passages in the report can be searched and retrieved from various dimensions such as country and year, which enhances the efficiency in searching for relevant information or materials.

Keywords: Information Aggregation; Information Unit; Information Organization; Fine-grained; Interactive Report

（收稿日期：2020-09-20）