• 医学信息技术 •

基础医学科研进展信息聚合平台构建*

郝志勇 庄永龙 张学工

（中国医学科学院基础医学研究所/北京协和医学院基础学院 北京 100005） （清华大学自动化系 北京 100084）

〔摘要〕 介绍基础医学科研进展信息聚合平台的构建，针对基础医学生物学相关内容，使用自动采集、粗分类、人工细分类、审核、批准发布的方式进行管理，提供丰富的展示方式和灵活安全的后台管理模式。同时可进行用户需求分析，并据此总结研究热点，为用户提供定制服务。

〔关键词〕 动态信息聚合；信息集成；垂直搜索；信息门户

The Construction of Scientific Research Progress Information Aggregation Platform for Basic Medicine *HAO Zhi-yong, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences/School of Basic Medicine, Peking Union Medical College, Beijing 100005, China; ZHUANG Yong-long, ZHANG Xue-gong, Department of Automation, Tsinghua University, Beijing 100084, China*

〔**Abstract**〕 The paper introduces the establishment of integrated information platform for basic medicine scientific research progress. Against related contents in basic biomedicine, which are managed by automatically collecting, coarse classification, artificial detail classification, audit and approval issued, providing rich information display ways, and safe back stage data management mode. Meanwhile the user demands can be analyzed so as to sum up hot research directions for the users' custom services.

〔**Keywords**〕 Dynamic information aggregation; Information integration; Vertical retrieval; Information portal

1 引言

信息聚合是指从不同的数据源汇集并分析相关信息、解决这些信息在语义方面的异构性，并提供基于数据源之间关系、业务过程的聚合等功能[1-4]。在大规模、动态、跨组织的网格环境下，需要提供一种松散耦合的信息聚合设施，它不仅能够实时地提供一致的信息视图、基于过程的信息集成，还需要能够适应动态变化的业务，例如业务的重构、新业务的扩展等[4-5]。传统数据集成的方法例如数据库管理系统等，不能适应这种不断变化的动态网格运行环境。面向服务的技术是应用于信息聚合的新趋势，信息聚合的操作被抽象为服务，并可以方便地进行重用和集成[6]。在提供信息聚合的动态性方面，关于领域应用中适应组织重构并基于变化的业务需求进行动态的信息聚合，从而提供服务给特定领域的用户等方面的研究还很少。本文从面向医学新闻及信息聚合入手，采用垂直信息搜索和信息聚合技术[7-9]，针对基础医学生物学相关内容，使用自动采集、粗分类、人工细分类、审核、批准发布的方式管理新闻资讯内容，并提供丰富的展示方式和灵活安全的后台管理模式，对设计一个支持动态信息聚合的服务网格具有参考作用[10-11]。

〔收稿日期〕 2010-04-23

〔作者简介〕 郝志勇，中级职称，发表论文 5 篇。

〔基金项目〕 科技部国家科技基础条件平台科学数据共享工程重大项目（项目编号：2005DKA32402）。

2 研发目标

国家人口与健康科学数据共享平台（原国家医

药卫生科学数据共享网）是国家科技基础条件平台科学数据共享工程的重大项目。基础医学科学数据中心是该共享网已启动的 6 个中心之一。项目的总体目标是建立一个物理上分布广泛、逻辑上高度统一的医药卫生科学数据管理与共享服务系统，为政府卫生决策、科技创新、医疗保健、人才培养、百姓健康和企业发展提供数据共享和信息服务。本项目是为基础医学科学数据中心提供相关的科研进展、文献报道、项目信息等。

整合全球基础医学相关数据源，提供丰富、主题明确、内容准确的基础医学信息资讯，保证信息内容的准确性、实时性和广泛性。可以根据用户访问频度和内容进行用户需求分析，并据此调整和修改新闻来源，总结热点方向，提供给用户针对性和导向性更明确的新闻资讯和新闻定制服务。构建快速、实时、自动、准确的基础医学新闻资讯门户，具体任务：（1）新闻资讯数据源采集：根据生物、医药、信息学经验，并采用专家推荐的方式确认目标准确、内容广泛、版权明晰的新闻来源；（2）新闻数据后台管理：针对新闻发布建立审核机制和流程，确保新闻内容准确、安全；（3）根据用户需求定制新闻资讯内容并实现 E-mail 自动提醒功能；（4）建立生物科学专有的词库，提供最方便快捷和精确的全文集成检索服务；（5）采用自主网页布局技术，构建专业级新闻发布门户网站；（6）严格内容筛选机制，保证发布新闻的安全性。系统具体架构，见图 1。

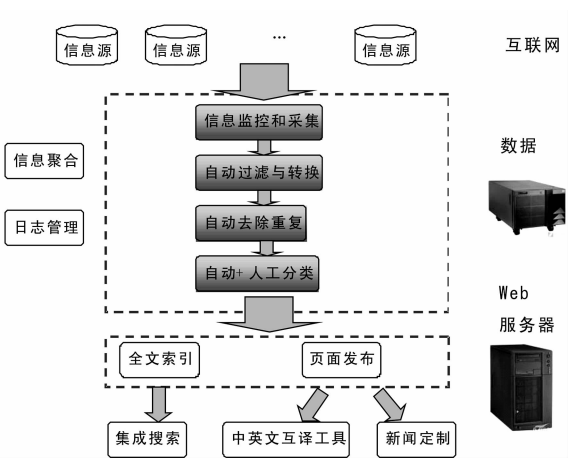

图 1 新闻聚合平台系统架构

3 技术路线

3.1 确认信息源

根据经验和专家推荐等多种方式确认信息源。代表性英文网站：Science Daily（http://www.sciencedaily.com），Nature（http://www.nature.com/），Biology News（http://www.biologynews.net/），BioArray News（http://www.bioarraynews.com/issues/），GenomeWeb Daily News（http://www.genomeweb.com/）。代表性中文网站：政府机关——科技部网站（http://www.most.gov.cn/），973 计划（http://www.973.gov.cn/），863 计划（http://www.863.gov.cn）；研究院所——中国生物技术信息网（http://www.biotech.org.cn），中国作物种质信息网（http://icgr.caas.net.cn/）；期刊杂志——科学网（http://www.sciencenet.cn/）。

3.2 信息自动聚合

基于垂直信息搜索和信息聚合技术，针对基础医学相关新闻资讯，建立自动下载分类系统[12-16]。管理人员可根据需要方便地指定需要监控采集的目标站点或频道，并设定监控更新的时间周期，包括定点执行、更新隔离，还可设置为自动轮转不间断运行。具备先进高效的采集技术和策略，采用多线程并发搜索技术和智能更新策略，每次仅采集最新更新过的网页，确保全过程合理高效。采集内容精准，多语言支持，适用性强，不仅能够采集常见的静态网页，还可以采集多种特殊形式的网页，包括：通过引入日期变量，可精确定位带有日期特征的栏目及其页面；引入页码变量，可有效采集基于数据库发布的多页信息；预设用户名和密码，可采集需要授权认证的站点或频道；引入模板，可精确定位采集页面局部特定内容等等。将所有页面转为 UTF8 编码，实现多语种网页的采集、存储和管理。可采集网页元数据和多媒体内容，可完整地识别并记录每个网页的详细原数据信息，包括网页名称、大小、日期、标题、文字内容等，网页中的图片和

表格信息可同时被采集。

3.3 自动过滤、分类和排重等智能信息处理功能

垃圾信息过滤，可对网页进行内容分析和过滤，自动去除广告、版权、栏目等信息，精确获取内容主体。基于内容自动排重，采用的排重技术不是简单的规则判断，而是利用内容的相似性进行判重，相似阈值可调，准确性高，不会因为标题或内容的少许变化产生漏判，即使改换标题系统也会正确判定。被检出的重复网页不会被及时清除，也可以作为主体网页的相似或者推荐网页进行显示，提供给用户参考。采用全文搜索引擎，基于近 50 万个生物医药专业词汇，可对网页进行无需人工干预的自动预分类，且准确率达到实用要求。机器采集分类的内容，必须由内容审核员审核通过之后，才能最终呈现到网站上。

3.4 内容分类分层扩展管理

对信息采用无限层级分类管理，满足最细节的分类需求，见图 2。采用新闻标签管理，各层级、各内容实体实现横向关联。

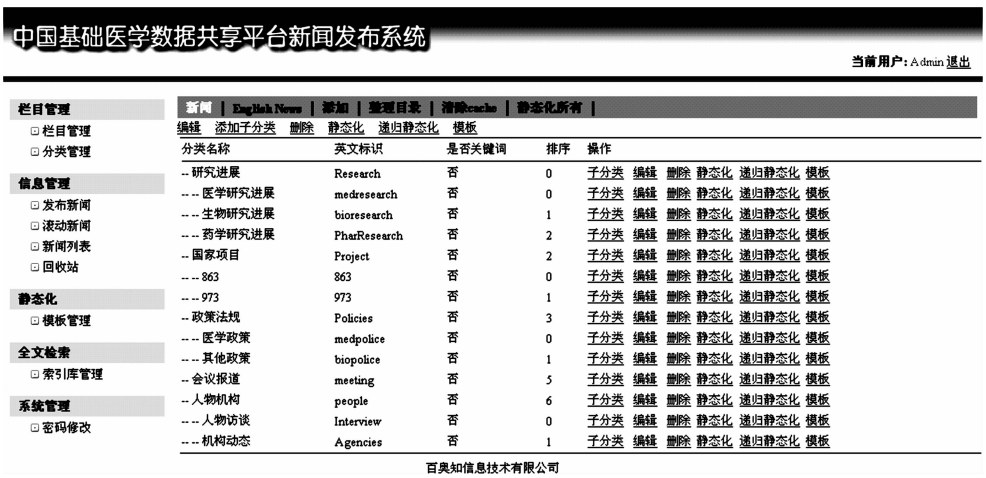

图 2 无限层级分类管理

3.5 灵活便捷的信息发布和检索

系统采集的信息可及时通过 Web 发布模块实时发布，界面以网页的形式展现，仅使用浏览器就可以查看和检索信息，方便易用。系统提供信息分类导航和检索功能，对于发布的信息，用户既可以查阅本地数据库中经过自动过滤的内容，也可以对照查阅原链接网页。系统支持自动发布和人工发布两种方式，在自动发布方式下，采集到的网页将自动发布到网站上；在人工发布方式下，采集的网页需要用户选择才能发布到网站上。可发布专题内容，支持自定制专题，用户可以通过定义关键词规则来建立专题，发布后专题中包含所采集到的满足条件的信息，方便用户跟踪特定主题的内容等。

3.6 丰富的页面布局与展现

灵活的网站布局，突破传统网站的固定行列布局，网站内容可以随意排布。无限扩展的表现形式，丰富的网站显示元素，如滚动图片新闻、横条图片新闻、标签页，并且能按照需要随意扩展。网站风格无限可定制，采用国际流行的 Web2.0 技术，页面主体结构为 DIV + CSS，风格定制非常灵活。

4 结语

网络信息时代，用户最大的需求在于信息的收集与管理，新闻信息聚合及面向特定用户群的搜索服务在网络资源指数级膨胀时期变得尤为重要[13-14]。“150 法则”指出当用户面对的信息超过

150 条时，用户就会感到困难。如何在最短的时间内以最快的方式将网络上信息聚合、分类然后再发布并有针对性地呈现给用户，成为各综合性信息网站发布平台研究探索的方向[2]。本文在中心前期工作成果和架构的基础上，进一步深入挖掘，通过垂直信息搜索和信息聚合技术，实现新闻信息自动聚合、自动过滤、分类和排重等智能信息处理、新闻内容分类分层扩展式管理、信息发布和检索、丰富的页面布局与展现等功能，为新闻信息聚合平台的建设与研究提供参考。

参考文献

1 姜恩波. 基于信息聚合的服务与技术〔J〕. 现代图书情报技术, 2007, (4): 32 - 34.

2 刘俊熙. 网络信息整合和检索功能的互动效应〔J〕. 现代图书情报技术, 2003, (S1): 135 - 137.

3 张会娥, 张智雄, 林颖, 等. 基于 RSS 的科技信息聚合系统的设计和实现〔J〕. 现代图书情报技术, 2005, (7): 60 - 63.

4 钱爱兵. 基于 RSS 的 Web 新闻主题聚合系统的设计与实现〔J〕. 现代图书情报技术, 2007, (4): 56 - 61.

5 陈凌晖. 基于 RSS 技术的信息门户个性化信息服务理念与实现〔J〕. 现代图书情报技术, 2007, (1): 33 - 36.

6 潘望, 朱宏明. 垂直搜索在个性化 Web 搜索中的应用〔J〕. 科技信息, 2008, (36): 89 - 90.

7 张秀虎. 浅析新闻采集程序的技术核心〔J〕. 中国教育信息化, 2007, (2): 55 - 58.

8 左敬龙, 余桂兰. RSS 新闻采集器的设计与实现〔J〕. 茂名学院学报, 2009, 19 (4): 39 - 41.

9 郝继英, 任慧玲, 王萍, 等. 基于概念图组织的中国艾滋病信息门户建设方案〔J〕. 现代图书情报技术, 2006, (12): 21 - 24.

10 殷蜀梅, 张智雄, 吴振新. 一种从医学文本中实现自动关键词抽取和筛选的技术方法〔J〕. 现代图书情报技术, 2008, (8): 31 - 36.

11 罗立群, 张慰, 陈金鑫. 基础教育黄页网站自动生成系统的设计与实现〔J〕. 现代图书情报技术, 2007, (8): 80 - 83.

12 许鑫, 黄仲清. 垂直搜索引擎应用中的若干策略探讨——以 12580 餐饮垂直搜索为例〔J〕. 现代图书情报技术, 2009, (2): 62 - 70.

13 吴浩东, 刘强. 一种信息门户中基于本体的信息查询模型〔J〕. 计算机工程, 2006, (16): 46 - 48.

14 赵艳丽, 李争艳. JSP 新闻发布系统〔J〕. 电脑知识与技术, 2007, (13): 175 - 176.

15 Hansen M, Madnick S, Siegel M. Data Integration Using Web Services〔C〕// Bressan S. ed. Proceedings of the VLDB 2002 Workshop Efficiency and Effectiveness of XML Tools and Techniques and Data Integration Over the Web, Hong Kong: Springer - Verlag, 2003: 165 - 182.

16 Sheth A, Larson J. Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases〔J〕. ACM Computing Surveys, 1990, 22 (3): 183 - 236.

网络参考文献著录规范

依据中华人民共和国国家标准《文后参考文献著录规则》(GB/T7714 - 2005)，网络参考文献著录格式如下：主要责任者. 题名:其它题名信息 [文献类型标志/文献载体标志]. 出版地:出版者,出版年(更新或修改日期)[引用日期]. 获取和访问路径. 。请各位作者严格按照国家标准著录网络参考文献。

《医学信息学杂志》编辑部