文献与数据学报
第1卷　第2期　　2019年6月

一种不依赖用户行为数据的科研文献推送系统*

虞哲英[1,2]　关　贝[1]　昝道广[3]　吕荫润[1,2]　毕丽阳[4]　王永吉[1,2,5]
（1. 中国科学院软件研究所协同创新中心，北京　100190；
2. 中国科学院大学，北京　100049；
3. 山东科技大学，青岛　266590；
4. 北京理工大学，北京　100081；
5. 中国科学院软件研究所计算机科学国家重点实验室，北京　100190）

摘要：［**目的**］实现一种智能推送科研文献的服务架构，解决在缺少用户行为数据时主动发起的文献精准推送问题。［**方法**］研究分析典型的推送系统和现存的文献情报服务，结合数据挖掘、数据存储、推送算法与可视化分析技术设计科研文献精准推送系统，并提出实现手段。［**结果**］系统整合了期刊文献录入、网络数据爬取、计算匹配学者、推广效果可视化和策略调整功能，期刊管理者可借助系统实现智能文献推送并发掘潜在读者。［**结论**］本文系统充分利用学术数据库公开的论文属性数据及作者属性数据，规避了科研文献个性化推荐系统常见的冷启动问题，为诸如期刊网站主动推送论文以吸引读者等应用场景提供支持。

关键词：精准推送　架构设计　数据挖掘　主题提取　用户画像
分类号：TP391.1
DOI：10.31193/SSAP.J.ISSN.2096-6695.2019.02.07

* 本文系科技部国家重点研发计划重点专项：“大数据驱动的中医智能辅助诊断服务系统”之子课题“多模态异构中医药大数据高效获取与资源库建设”（课题编号：2017YFB1002301）研究成果之一。

［作者简介］虞哲英，男，中国科学院大学人工智能学院，学生，本科，研究方向为人工智能、软件工程，Email：j_y_1@sina.com；关贝，男，中国科学院软件研究所，助理研究员，博士，研究方向为人工智能方法和大数据分析方法、网络安全分析技术、操作系统虚拟化技术和安全操作系统，Email：guanbei@iscas.ac.cn（通讯作者）；昝道广，男，山东科技大学，本科生，研究方向：自然语言处理，Email：zandaoguang@163.com；吕荫润，男，中国科学院软件研究所协同创新中心，博士生，研究方向：析取规划、运筹优化、自然语言处理，Email：yinrun@nfs.iscas.ac.cn；毕丽阳，女，北京理工大学，硕士生，研究方向：推荐系统，Email：biliyang@gmail.com；王永吉，男，中国科学院软件研究所，博士生导师，二级研究员，博士，研究方向为人工智能、大数据、云计算、软件工程、虚拟化技术、隐蔽信道、实时系统等，Email：ywang@itechs.iscas.ac.cn（通讯作者）。

0 引 言

互联网时代，学者借助网络就可以获取科研文献，大型的学术数据库和专门的搜索引擎也提供了丰富的检索功能。但通过关键词和多重条件筛选的方式来寻找真正亟需、与自己聚焦的领域贴合的高质量文献，仍是一件耗时耗力的工作。

本文在分析对比现有的文献搜索和推送服务的基础上，充分挖掘文献主题摘要数据、作者属性数据及各类学术合作关系数据的潜在价值，综合运用数据挖掘、用户画像、主题提取、智能推送等前沿技术，提出了不依赖用户行为数据的科研文献精准推送系统的架构。系统避开了一般个性化推荐系统中的冷启动问题，从另一角度尝试进行科研文献的精准推送，提升了知识的传播效率。

1 推送技术的应用现状

1.1 典型的推送系统应用

1.1.1 电子商务平台推送系统

电子商务平台是推送系统在互联网上最热门的应用领域之一，人们熟知的亚马逊、淘宝、京东等电商平台均拥有数亿的注册用户和千万级的日活跃度。这类电商系统主要采用基于用户的协同过滤算法和基于商品的协同过滤算法，通过记录和分析用户对商品、卖家的评分信息，找出与当前用户在评价行为上相似度高的用户，并将后者看好的商品推送给当前用户[1]。协同过滤技术在积累了大量用户评价数据的电商系统中展现了不错的准确性，但对于未积累足够数据或者新类型商品的场景存在冷启动和稀疏的问题。

1.1.2 个性化资讯服务

互联网时代人们面临信息爆炸和知识过载，谁能够因人而异的提供个性化的信息服务，谁就能赢得用户的青睐，新闻资讯端的今日头条、知识问答端的知乎，都是典型的成功应用。以今日头条为例，通过对用户兴趣和触媒行为做深度分析挖掘，用“兴趣图谱”进行用户画像，形成一套庞大的用户 - 内容标签体系，从而为资讯和用户进行适当的匹配，做到“千人千面”的信息内容推送。今日头条的个性化推荐技术除了也使用和电商类似的协同过滤算法，还使用基于内容的推荐方式，通过机器学习算法对新闻内容进行刻画，然后利用用户的正负反馈建立用户和新闻标签之间的联系，同样依赖对用户行为数据的大规模收集。

1.1.3 音视频推送引擎

听音乐、看视频已成为很多人的娱乐方式，很多音视频网站引入推送系统为用户提供其感兴趣的音视频，吸引更多的用户访问。全球最大的视频网站 YouTube 基于上传视频者标记的标签进行推送；国内的优酷、腾讯、爱奇艺等网站在标签的基础上采用召回 + 排序，召回值得推送的视频按关注度进行排序，将关注度高的视频推送给目标用户。这些视频网站使用的推送技术均依赖用户行为数据。

1.2 科研文献推送服务现状

1.2.1 数字图书馆服务

图书馆是人类知识文化资源的集中地，数字图书馆则是在互联网环境下应运而生的知识中心，其以数字技术处理和存储文字、图片乃至音频视频，将馆藏通过网络实现信息资源共享，为用户提供更方便、快捷、高水平的信息化服务。现有数字图书馆的推送服务包含三种类型：根据用户感兴趣的文献推荐馆藏中与其相似度高的文献，即基于内容的推荐；分析用户历史偏好将相似用户感兴趣的文献相互推荐，即基于协同过滤的推荐；以及将两者特点分析融合的混合推荐。但是与百度学术、知网等学术数据库相比，独立的数字图书馆在资源的种类和规模上都难以比拟，未来发展正受到挑战[2,3]。

1.2.2 学术数据库服务

上世纪90年代，主流商业期刊出版社、学术组织和发行中间商构建学术期刊数据库以集中电子学术资源[4]。作为全球最大的全文数据库，ScienceDirect将Elsevier出版的2500多种期刊和11000种图书全部数字化，通过网络提供服务。而中国知网作为最大的中文期刊全文数据库，收录了国内8200多种重要期刊，全文文献总量达2200万篇[5]。为解决庞大文献数量给用户发现所需论文带来的困难，知网提供了丰富的搜索和推送功能。用户在一篇论文的详情页面可以看到内容相似的论文，以及读过该论文的其他读者还对哪些文献感兴趣。在积累了一定的用户搜索数据后，知网也会在“我的CNKI”页面为用户推送一些新发表的文献。

1.2.3 期刊网站

为了提高期刊的影响力和显示度，有很多文章从期刊内容、期刊宣传、期刊出版、人力资源、期刊网站等多个方面给出了经验和建议。期刊网站作为期刊的数字化传播平台，除了可以为读者提供线上阅览渠道，还可以发布学术活动或提供形式各异的附加服务，以期扩展期刊影响力[6-8]。这些网站发挥了期刊在内容和专业方面的优势，其中很多也提供了开放存取服务以提升科学研究的效率及公共利用程度。主流期刊网站采用订阅的方式定期将新刊文章推送至用户邮箱，但这类推送通常不会依据对象特征进行个性化定制，或者只是简单的让用户自己从若干分类中进行选择，最后用户收到的推送文章与自己兴趣匹配的几率不高。

1.3 现状分析

电商、资讯和音视频等商用领域的个性化推荐系统多以协同过滤推荐技术为基础。这类系统一般都提供用户对推荐对象的评价功能，通过收集和分析大量的用户评价信息，形成比较可靠的推送结果。很多现有的科研文献推送服务虽然也采用协同过滤推荐，但因为不易获得评价信息，加剧了稀疏性和冷启动问题，难以提供令用户满意的推送结果。

本文提出的科研文献推送系统，是基于主动抓取的属性数据，分析科研用户的研究领域，从而可以在不借助用户行为数据的情况下进行精准的期刊文献推送，帮助期刊网站挖掘潜在读者，提升科研文献服务的效果。

2 架构设计

科研文献精准推送系统主要面向科研文献内容的管理者和推广者，提供针对具体文献的精准

推送服务以及推送效果的统计分析支撑。精准推送系统的架构设计以采集的属性数据为基础，以推荐算法为主要支撑，并以提供满足需求的应用服务为目标。

本文提出的科研文献推送系统采用分层和模块化的架构设计，力求各个功能模块高内聚低耦合、更易维护和拓展，由于数据采集任务比较耗时，系统使用独立的任务管理模块在后台自动发起采集，推荐算法、统计分析等任务直接使用数据存储中已有数据，从而和采集任务互不干扰、并行展开，以保证系统的性能和稳定性。

科研文献推送系统的整体架构如图 1 所示，分为四个层次：数据采集层、数据存储层、推送算法层和应用服务层。

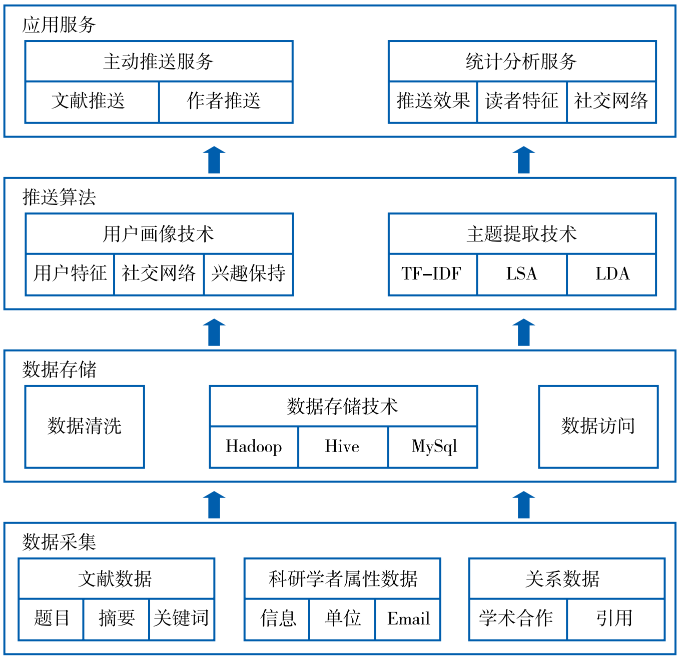

图 1　系统总体架构

3　数据采集

数据采集层的职责是收集挖掘系统所需信息。使用系统录入和网络爬虫两种方式，获取必要的文献数据、科研学者属性数据和学术合作关系数据，提取并转化为结构化数据输入到数据存储层。数据采集的常用方法包括人工采集和网络爬虫两种。如果掌握了采集目标网页的数据组织结构，网络爬虫技术就可以连续、自动地从目标数据来源网站抓取所需数据项，并储存为数据库中的结构化数据。在科研文献精准推送的应用场景中，目标网站是知网、Web of Science 等发展成熟的学术数据库网站，这些网站数据信息全面，且页面结构比较稳定，少有大幅变动，很适合爬虫抓取。

科研文献精准推送系统所需的数据信息，按照采集的数据类型不同，可以分为文献数据（Literature Data）、科研学者属性数据（Attribute Data of Scientific Research Scholars）以及关系数据

（Relational Data）三类。

科研文献精准推送系统的网络爬虫模块使用的是 geckodriver 技术，依托于 Mozilla 的服务器，以此实现远程网页内容的爬取。Geckodriver 依靠 webdriver crate 提供 HTTPD 服务，完成繁重的 WebDriver 工作。同时，Geckodriver 还可以作为 WebDriver 和 Marionette 中间的代理，将 WebDriver 命令、错误和响应转换为 Marrionette 协议。通过此技术，本系统数据采集模块可以根据设定的规则进行科研文献的爬取。

3.1　数据采集架构

本系统首先输入要推送的文章，然后使用 geckodriver[9] 模拟登陆 web of science 网站，之后模拟输入论文的关键字，爬取本页内容并存入数据库，接着判断是否有下一页，如果为是，则继续爬取本页内容，如果为否，则结束爬虫。

在爬取内容阶段，主要分为三大部分，文献数据采集、科研学者属性数据采集以及关系数据采集。文献数据采集架构如图 2 所示。

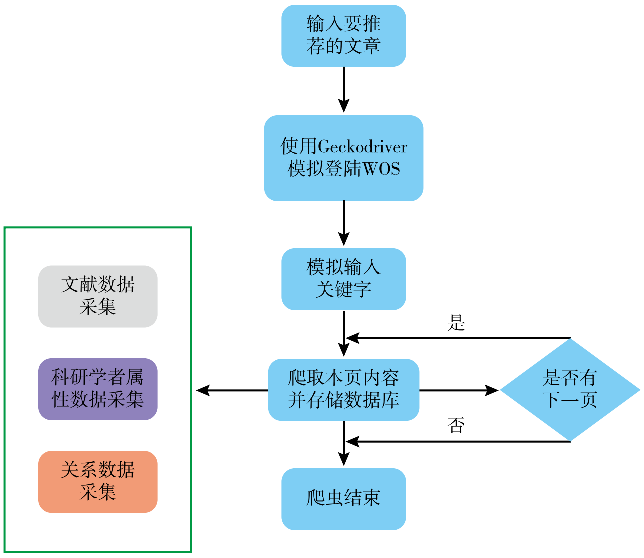

图 2　文献数据采集架构

3.2　文献数据采集

首先通过 geckodriver 按照规则将 web of science 数据库中的文献采集下来，然后利用 TF-IDF[10] 算法求解被推送文献和爬取的文献相似度，最后根据相似度进行精准推送，以满足期刊的推广需求。

此部分主要采集文献的题目、卷、期、页、文献号、DOI、出版年、文献类型、摘要以及关键词等信息。

针对文献数据采集，按照特定规则爬取文献并存入数据库后，可以采用级联的思想进一步采集数据，例如经过自然语言处理之后，得出 A 文献的作者极有可能对待推送的文献感兴趣，由此级联到 A 文献的参考文献对应的作者，以及 A 文献作者发表的其他文献的参考文献对应的作者等。

3.3 科研学者属性数据采集

伴随文献数据的采集，首先把科研学者的信息按照规则采集下来，然后利用邮箱匹配算法对其进行作者和邮箱匹配［11］。将采集下来的作者邮箱存入数据库，按照推送优先级发送邮件。

此部分主要采集作者的姓名、研究方向关键词、邮政地址、电子邮箱、所属机构等属性数据。

针对科研学者属性数据的采集，本文系统遵循的原则是：如果此作者没有邮箱，则将作者和邮箱放弃存储，如果此作者存在邮箱，则将作者和邮箱存储到数据库。下一步计划如果该作者没有邮箱，则查找此作者的其他文章中是否存在邮箱，如果有，则存储，反之，则彻底放弃（如图 3）。系统使用机构名 + 姓名来标识作者，从而降低因重名导致采集错误的可能性。

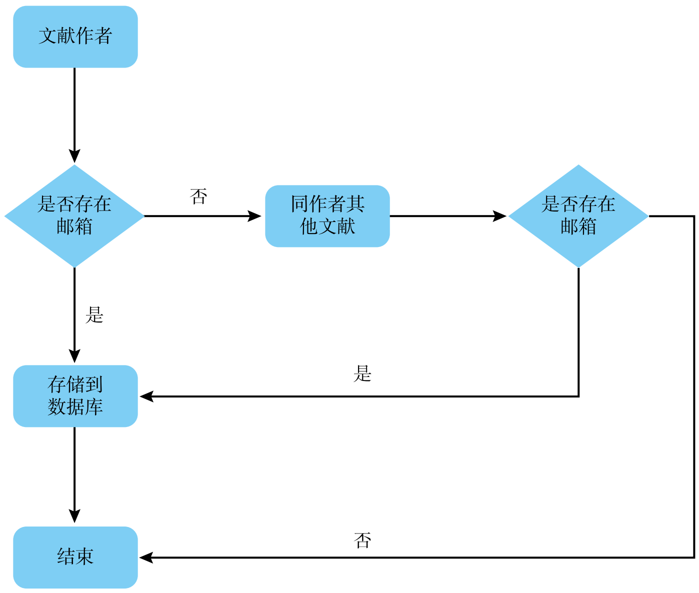

图 3　文献数据采集改进策略

3.4 关系数据采集

为了提升科研文献精准推送系统的精确度和惊喜度，系统还会采集作者和文献的关系数据，作为社交网络分析模块的输入。关系数据主要包括文献间的引用关系、研究人员的合作关系以及研究人员和组织机构的所属关系。其中，文献引用关系和学者合作关系从文献页面爬取，每篇文献存为一条记录，并用两个字段分别保存其引用的所有文献 ID 和所有作者 ID，ID 间以分隔符分割。人员机构所属关系从作者页面爬取，保存在人员属性表。

系统通过对上述数据的分析，可以发现研究人员之间的社群关系，从而以不同于文本相似的另一角度对推荐结果进行修正。

4　数据存储

数据存储层的目标是要把采集层获取的数据进行统一格式化处理，根据实际需要采用一

定的数据存储架构进行存储，并提供后续数据分析和加工的访问接口。数据存储层主要包含几个功能模块：数据清洗，数据存储，数据访问。清洗模块将多源异构数据统一格式化处理，保证数据质量。存储模块采用 Hadoop、Hive 和 MySQL 结合的分布式文件系统进行数据永久化存储，从可靠性、易用性和性能三个方面满足用户需求。访问模块作为数据存储和数据服务的纽带，使用 RESTful 架构来实现系统 Web 服务对存储数据的访问请求，减少系统的耦合度。

4.1　数据清洗

数据清洗的主要目的是对获取到的多源异构数据进行整理，并统一格式化处理，使得在数据的整个生命周期中系统能够提供高质量的数据服务。本文研究的数据清洗有两个主要过程。首先对获取的数据进行过滤筛选，去除无效数据，保持数据的完整性和有效性。爬虫获取到的数据主要包括文献数据和作者数据，文献数据格式包括文献的相关属性，如中英文标题、中英文摘要、中英文关键词、作者、发表年份、卷期号、页码、文献号等，部分文献包括全文信息等，作者数据主要包括作者姓名、单位、Email 地址等。本研究的方式是通过全文匹配文章摘要进行推送，因此对于没有摘要的文献数据，可直接进行过滤。对于作者数据，本工作是为了获取作者的 Email 地址，没有 Email 的作者信息就是无效数据，需要清洗掉。其次，数据清洗需要统一数据标准。多源异构的数据存在格式不统一、符号混用等问题，需要在存储之前进行标准化处理。例如针对日期格式问题，系统设计了专门的日期分析模块，采用十余种日期格式转换公式对源数据进行匹配，解码为数值型日期进行存储，覆盖了学术数据库中可能出现的绝大多数日期数据。

4.2　数据存储

数据存储层将爬虫数据进行永久化存储，提供给系统进行分析使用，以及共享给第三方进行使用，并进行数据的定期更新。

考虑到数据存储的可靠性、易用性和系统性能，本文系统选择 Hadoop、Hive 和 MySQL 结合的分布式文件系统来存储数据。科研文献推送系统预计要爬取十万、百万级的文献数据，每篇文献包含数十条属性项目，数据量庞大。Hadoops 分布式文件系统 HDFS 能提供高吞吐量的数据访问，支持 PB 级的数据存储，非常适合大规模数据集上的应用，可以满足系统的存储和访问需求。另外，HDFS 具有高容错性，可将廉价闲置机器的存储空间利用起来提供存储服务，降低了系统的硬件门槛。系统执行推荐算法的第一步是读取文献数据，由于算法复杂，计算步骤所需的时间消耗难以避免，我们需要尽可能提升访问步骤的性能。Hive 是基于 Hadoop 的数据仓库工具，可将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能，将 SQL 语句转化为 MapReduce 任务来执行，从而大幅提高数据的访问效率[12]。Hive 将元数据存储在数据库中，如 MySQL 等，Hive 元数据包括表的名字，表的列和分区等属性。系统将清洗后的文献数据和作者数据按照属性存储到不同服务器的结构化数据库 MySQL 中，使用 Sqoop 工具可在关系型数据和 Hadoop 的 HDFS 之间建立数据传输桥梁，如图 4。该存储架构可实现海量数据存储的可靠性，满足用户访问数据的性能需求，同时也能降低构建专属数据中心的成本。

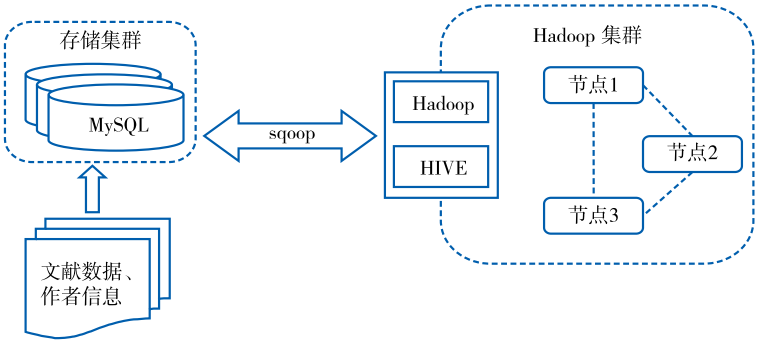

图 4　系统数据存储架构示意图

4.3　数据访问

数据访问层是连接数据存储和数据服务的纽带，数据访问层要能在系统的不同层之间传递数据，同时要能满足重用性、并发性、可扩展性等需求。科研文献推送系统的数据采集、数据清洗和推荐算法模块各自均具有一定的复杂性，考虑到增加采集数据源、补充格式化规则、优化推荐算法等可能存在的后续需求，必须避免模块之间的依赖带来更大的维护和扩展难度。本文系统使用 RESTful 架构来实现 Web 服务对文献数据的访问[13]，RESTful 是指 Representational State Transfer，即表现层状态转化，它是一组架构约束条件和原则，主要有两个特点。首先，RESTful 要求数据服务中客户端和服务器之间的交互在请求之间是无状态的，客户端到服务端的每个请求都必须包含理解请求所必需的信息，无状态请求可被任何可用数据服务器回答，适合云计算之类的分布式架构环境，另外客户端也可以进行缓存数据以改进性能。其次，RESTful 是分层系统，组件无法了解与之交互的中间层以外的组件，通过限制系统知识在组件之间，可降低系统的复杂度，减少系统的耦合度，为系统的优化和扩展提供良好环境。

本文系统将文献数据和作者数据等，抽象成数据服务对象资源，用数据抽象层来提供，客户端基于 HTTP 协议和 RESTful 服务协议，使用 GET/POST 等通用 Web 请求来访问所需数据资源，服务器通过相应的应用进行授权和分发，对客户端的数据请求进行响应，如图 5 所示。基于 RESTful 的服务架构可实现数据的共享和独立并发访问。

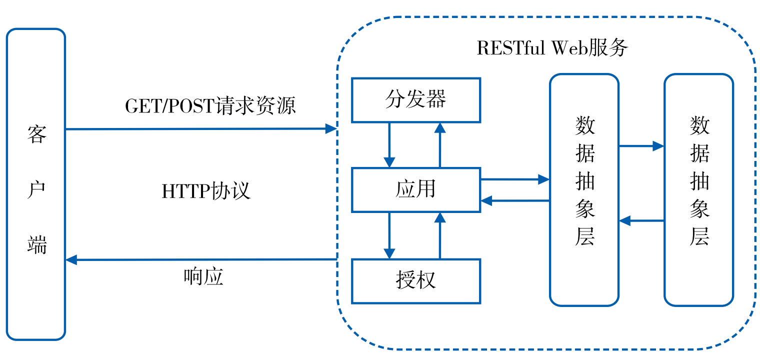

图 5　系统 RESTful 数据访问协议示意图

5 推送算法

推送算法层是文献推送系统的技术核心部分，由用户画像、主题提取和精准推送三个模块构成。用户画像模块依据爬取的科研学者属性信息分析期刊潜在用户的特征，主题提取模块从待推送和爬取的文献著作文本中分析主题词，精准推送模块根据前两个模块的分析结果计算最佳的推送匹配对象。

5.1 用户画像

用户画像是根据用户人口统计学信息、社交关系、偏好习惯和消费行为等信息而抽象出来的标签化画像，是精准推送系统的基础。科研文献精准推送系统的目标用户是科研学者，依据采集到的学者属性信息、发表文献信息、学术关系信息以及系统运行中产生的反馈信息，勾勒出科研学者用户的需求全貌[14]。

5.1.1 用户特征模型

用户特征模型采用数据挖掘的技术，描述用户的个人属性、偏好信息及历史行为记录，从而发现并预测用户的潜在需求。本系统以学者发表过的文献主题词表征其静态的科研领域偏好，并以学者对推送结果的反馈情况表征动态的行为属性，确定了如 $\alpha \cdot I^s + (1-\alpha) \cdot I^d$ 的特征模型[15]。通过主动地从学术数据库爬取论文的相关字段，规避了系统的冷启动问题。

5.1.2 社交网络关系

通过分析学者所属的组织机构、文献之间的引用关系以及学术库提供的著者关注图谱，系统可利用学者的社交网络特征来辅助用户兴趣模型的构建[16]。社交网络用单向关注的形式进行有向图计算，如下式：

$$P_{ui} = \sum_{v \in F(u)} w_{uv} r_{vi} \tag{1}$$

其中，F(u) 是学者 u 的全部关注学者，r_{vi} 是学者 v 感兴趣的文献 i，w_{uv} 代表了学者 u 和 v 之间的关注指数或相似度。

5.1.3 兴趣保持模型

在一段时间内，用户可能对某类文献感兴趣，也可能会对某类文献失去兴趣。这种兴趣动态变化的现象和认知记忆的遗忘机制类似。将时间对用户短期兴趣变迁的影响加入考虑之中，假设其发表文献中的某个关键词或主题词的累积次数可以反映某一时期内的兴趣，则兴趣保持模型可以用以下公式表示：

$$CI\left(t_{e(k)}(i), m\right) = \sum_{j=1}^{m} y_{t_{e(k)}(i),j} \tag{2}$$

式中，$y_{t_{e(k)}(i),j}$ 表示 j 时间段内，用户发表的关于领域本体 e（k）的主题 $t_{e(k)}(i)$ 的文献数量。兴趣保持模型可以计算用户的保持兴趣，预测当前最感兴趣的研究课题，从而使系统的推送更具时效性。

5.2 主题提取

在用户建模和文献建模的过程中，都需要从学者发表文献或待推送文献当中提取能够表达文

献特征的词语，形成特征向量。可以取得的文献标识包括作者、单位、标题、关键词、摘要、正文、分类号、刊名等，系统分别使用潜在语义分析（Latent Semantic Analysis，LSA）和潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）构建文献主题模型[17]。

5.2.1　潜在语义分析

LSA 用向量空间模型处理文献内容信息，将其转换为一组关键词集合。对于摘要、标题这类未分词的语言信息，系统采用自然语言抽取技术先转换为关键词信息，再将全部关键词按权重排名，得到最终的关键词向量，以供与用户特征进行匹配，形成推送。关键词向量的表达如下式：

$$d_i = \{(e_1, w_1), (e_2, w_2), \ldots\} \tag{3}$$

其中 e_i 代表一个关键词，w_i 代表该关键词的相应权重。对于权值的计算，适合本文系统的做法是采用 TF-IDF（Term Frequency-Inverse Document Frequency）公式：

$$w_i = \frac{TF(e_i)}{\log DF(e_i)} \tag{4}$$

TF-IDF 算法注重关键词共现，对于专业词汇较多的科研文献效果显著。得到当前文献的关键词向量后，与带匹配文献的相似度可以通过两组关键词向量之间的余弦相似度计算，见下式：

$$\mathrm{sim}(d_i, d_j) = \frac{d_i \cdot d_j}{\sqrt{\|d_i\|\|d_j\|}} \tag{5}$$

为了提升性能，避免在时间和硬件上的过多消耗，系统按照学科逻辑结构分类将文献进行分组计算，同时建立“关键词 - 文献”倒排表，只从关键词相似度较高的文献中进行匹配。

5.2.2　LDA 主题模型

LDA 是一种用来识别文本中潜藏的主题信息的非监督机器学习技术[18]。模型中每篇文档都是一个由 N_d 个单词构成的向量 $\overline{w_d}$，其中每个单词都属于数量为 V 的词典。一个 M 篇文档的集合定义为 $D = \{\overline{w_1}, \overline{w_2}, \ldots, \overline{w_M}\}$，LDA 假设在这个数据集中有一组潜在的主题 Z = $\{z_1, z_2, \ldots, z_T\}$，T 为预先设定的主题个数。每篇文档被看作是由 Z 中所有主题的分布向量，每个主题被看作词典 V 上的单词分布向量。LDA 的生成模型如图 6：

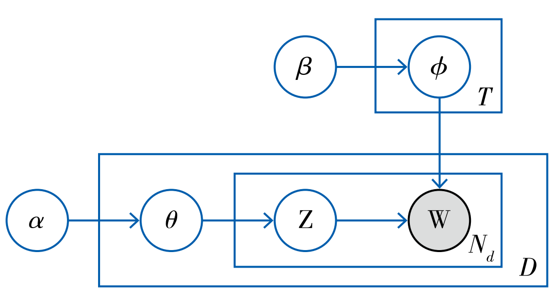

图 6　LDA 生成模型

图中，每篇文档与 T 个主题的多项式分布记为 θ，每个主题与 V 个单词的多项式分布记为 φ。θ 和 φ 分别有一个带超参数 α 和 β 的狄利克雷先验分布，通过学习这两个参数，可获取文档作者感兴趣的主题及每篇文档所涵盖的主题比例信息。

5.3　精准推送

系统启用初期，使用基于内容的推送算法，依照前两节描述的方法对用户和待推送文献建模，找出与学者所发表文献的主题最为相似的期刊文献进行推送。由于采用主动从学术数据库爬取学者信息的方式，规避了推送系统经常面临的冷启动问题，同时可以发掘期刊的潜在读者。

形成推送之后，系统会追踪目标邮件的浏览和点击行为，作为推送效果数据进行统计分析。当采集的学者信息数据和推送反馈数据达到一定规模后，可根据形成的学者社交网络和文献评价数据，使用协同过滤算法寻找学术关系和评价模式相近的用户群体，把群体中个体评价优秀的文献推送给其他用户。

6　应用服务

应用服务层是系统使用者可见的顶层功能，包括主动推送和统计分析两大模块。主动推送模块让期刊管理用户便捷地导入期刊文献数据，配置推送参数并执行推送操作；统计分析模块从多个维度直观呈现推送效果，帮助管理人员掌握用户特点并调整策略。

6.1　主动推送服务

论文精准推送系统强调知识服务的主动意识和用户意识。主动意识指期刊单位要发挥服务的主动性，依据挖掘到的数据和算法分析的结果，明确怎样的内容可以满足服务对象的需求，主动将高质量的内容提供给匹配的读者，使出版文献得到更广泛的传播和利用。用户意识指期刊单位应注重用户分析，积极收集用户属性信息和行为反馈数据，从多角度挖掘用户的特征和需求，重视用户体验，提供更便捷和智能的服务。

6.1.1　文献精准推送

本系统的主动推送服务以线下邮件的形式为主。科技期刊通过邮件将收录论文发送给读者的情况并不少见，但大多缺乏个性化匹配，论文和读者兴趣南辕北辙，导致推送效果不佳。精准推送系统针对每篇期刊论文，经过数据采集模块的学者挖掘和推送算法模块的相似度计算，选取匹配程度最高的若干学者形成潜在的受众列表，辅以匹配分值，展示给系统的期刊编辑用户。期刊编辑可以在此基础上进行人工筛选，并自定义邮件样式，执行邮件推送。

6.1.2　作者精准推送

在积累了一定数量的期刊文献后，系统可形成期刊的作者数据库。用与期刊文献的主题提取类似的方法，抽象出论文发表者的特征模型，通过推送算法寻找感兴趣的服务对象，以邮件形式推送作者的研究领域和历史著作。相较于单篇学术论文蕴含的信息，一个研究领域相近的高水平作者的研究工作能为服务对象带来更多帮助，更容易引起关注。

6.2　统计分析服务

作为系统直接使用者的期刊编辑人员，需要直观、实时、多角度地查看科研文献推送的反馈

数据统计，从而了解推送效果，把握服务对象的特点，支撑推送策略的调整。

6.2.1　推送效果分析

科研文献推送服务是一种主动发起的推送行为。从待推送内容的角度出发，推送效果可以按单篇文献、关键词、主题领域以及推送时段的指标进行多维统计（如图 7），判断一段时间内的热点方向；从推送行为的角度出发，可以分期刊编辑、邮件模板样式观察反馈效果，找出最佳的人工优化实践。

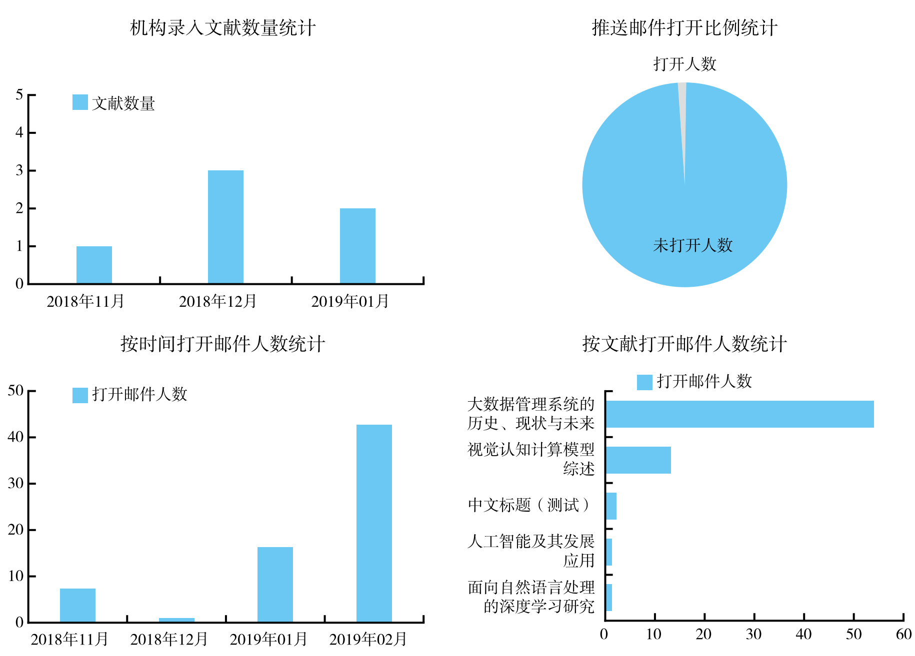

图 7　推送效果分析示意图

6.2.2　读者特征分析

主动推送的目的在于发掘和了解潜在读者。监测到打开邮件和点击推送内容链接的行为后，系统根据邮箱地址对应已采集并存储的科研工作者的所属机构、担任职位、所在地域、研究方向等属性特征，将其和推送文献的关键词、主题词挂钩统计，以多维图表的方式呈现，让期刊管理者直观地掌握对不同类型文献感兴趣的读者特点。

6.2.3　社交网络分析

前面提到，数据采集模块采集了诸如文献间引用、合作作者、作者所属机构等多类关系数据，经过社交网络算法的整理和计算，形成学术合作关系的社交网络。社交网络分析是对社交网络进行研究，为多结点之间的关系进行描述，并对其价值进行估量的一个工具。

7　结　语

本文旨在提出一种在缺少用户行为数据的场景中，仍能将科研文献准确地推送给对其感兴趣

的学术研究者的系统架构。在总结了现有典型的个性化推荐技术和文献情报服务的基础上，本文设计出分层模块化的系统架构，从数据采集、数据存储、推送算法和应用服务四个层面阐述数据挖掘和个性化推送技术在科研文献主动推广方面的应用。本文系统在推送算法的动态优化方面还不够完善，未来需要进一步评测验证系统的推广效用。

[参考文献]

[1] 黎超 . 基于大数据的电商个性化推荐系统分析 [J]. 商业经济研究 ,2019(2):69–72.

[2] BEEL J, GIPP B, LANGER S, et al. Research–paper recommender systems: a literature survey [J]. International Journal on Digital Libraries, 2016, 17(4):305–338.

[3] NASCIMENTO C, LAENDER A H F, DA SILVA A S, et al. A source independent framework for research paper recommendation [C] //NEWTON G. Proceedings of the 11th annual international ACM/IEEE joint conference on digital libraries. New York: ACM, 2011: 297–306.

[4] 王炎 . 数据挖掘技术下的个性化智能推荐系统设计 [J]. 微型电脑应用 ,2019,35(2):119–121.

[5] 刘瑾 .Web 系统 Selenium Web Driver 自动化测试框架搭建 [J]. 电子技术与软件工程 ,2017(21):171–172.

[6] 王维朗，吕赛英，游滨等 . 提升科技期刊国际显示度的途径与策略 . 中国科技期刊研究 ,2011,22(5):743–745.

[7] 王萍，甄志勇，徐状 . 中国科技期刊学术影响力的提高与建设 [C]//. 中国科学技术协会学会学术部 . 第六届中国科技期刊发展论坛论文集 ,2010:225–230.

[8] 彭桃英，周红梅 , 雷燕，等 . 网络环境下提高中国科技期刊影响力的途径 . 农业图书情报学刊 ,2013,25(8): 164–167.

[9] 于韬，王洪岩 . 基于 TF–IDF 算法的文本信息提取 [J]. 科技视界 ,2018(16):117–118.

[10] 赵瑞 . 邮址精确提取及邮件针对性发送系统开发 [D]. 浙江理工大学 ,2015.

[11] 张成博 . RDBMS 到 Hadoop 数据与 SQL 迁移的研究与实现 [D]. 华东理工大学 .2018.

[12] 张志，胡志勇 . RESTful 架构在 Web Service 中的应用 [J]. 计算机应用 ,2018,37(10):33–37.

[13] Gema Bello–Orgaz,Jason J. Jung,David Camacho. Social big data: Recent achievements and new challenges [J]. Information Fusion,2016,28.

[14] 熊回香，杨雪萍，高连花 . 基于用户兴趣主题模型的个性化推荐研究 [J]. 情报学报 ,2017,36(9):916–929.

[15] 黄继婷，陈建兵，陈平华 . 融合偏好度与网络结构的推荐算法 [J]. 计算机工程与应用 ,2019(10):1–10.

[16] 牛永洁，田成龙 . 融合多因素的 TFIDF 关键词提取算法研究 [J]. 计算机技术与发展 ,2019(7):1–5.

[17] XU Ke, ZHENG Xushen, CAI Yi, et al.Improving user recommendation by extracting social topics and interest topics of users in uni–directional social networks [J].Knowledge–Based Systems,2018,140:120–133.

[18] 裴超，肖诗斌，江敏 . 基于改进的 LDA 主题模型的微博用户聚类研究 [J]. 情报理论与实践 ,2016,39(3):135–139.

A Research Paper Recommendation System Independent of User Behavior Data

YU Zheying[1,2] GUAN Bei[1] ZAN Daoguang[3] LYU Yinrun[1,2]
BI Liyang[4] WANG Yongji[1,2,5]

(1. Collaborative Innovation Centre, Institute of Software Chinese Academy of Sciences, Beijing 100190,China; 2.University of Chinese Academy of Sciences, Beijing 100049,China; 3.Shandong University of Science and Technology, Qingdao 266590,China; 4.Beijing Institute of Technology, Beijing 100081,China; 5.State Key Laboratory of Computer Science, Institute of Software Chinese Academy of Sciences, Beijing 100190,China)

Abstract:［**Purpose/significance**］Software architecture for intelligent recommendation of research papers is implemented to solve the problem of precise pushing of papers initiated by the system in the absence of user behavior data.［**Method/process**］This paper studies and analyses the typical recommendation system and the existing document and information services, designs the precise pushing system of research papers by combining data mining, data storage, push algorithm and visual analysis technology, and puts forward the means to realize it.［**Result**］The system integrates the functions of journal paper input, network data crawling, calculating matching scholars, and visualization of promotion effect and strategy adjustment. Periodical managers can use the system to realize intelligent paper delivery and discover potential readers.［**Conclusion**］This system makes full use of the publicly available paper attribute data and author attribute data in academic databases, avoids the common cold start problem of personalized recommendation system in the field of research papers, and provides support for application scenarios such as journal websites actively delivering papers to attract readers.

Keywords: Precise recommendation; Architecture design; Data mining; Topic extraction; User profile

（本文责编：王秀玲）