·应用研究·

基于群组平台知识圈的精准信息推荐

王 峰[1] 林丽珊[2] 刘 毅[1]
（1. 中国科学院武汉文献情报中心，湖北 武汉 430071；
2. 中国科学院广州能源研究所，广东 广州 510000）

〔摘 要〕为了实现面向科研群组的精准信息推荐，满足科研人员更方便快捷的获取信息的需求。本文在汇聚知识资源形成群组平台知识圈的基础上，构建群组兴趣模型，计算推荐信息与群组兴趣的相似度，实现了科研资讯和领域监测快报的自动精确推荐，为科研群组平台的可持续发展提供了良好技术支持。

〔关键词〕群组平台；知识圈；精准信息推荐
DOI: 10.3969/j.issn.1008-0821.2018.07.011
〔中图分类号〕G250.76 〔文献标识码〕A 〔文章编号〕1008-0821（2018）07-0074-07

Accurate Information Recommendation Based on Subject Groups Integration Platform Knowledge Circle

Wang Feng[1] Lin Lishan[2] Liu Yi[1]
（1. Wuhan Documentation and Information Center of the Chinese Academy of Sciences，Wuhan 430071，China；
2. Guangzhou Institute of Energy Conversion，Chinese Academy of Sciences，Guangzhou 510650，China）

〔**Abstract**〕The paper aimed to achieve accurate information recommendation for subject groups，meeting the demand of researchers to obtain information more conveniently and quickly. First，it gathered knowledge resources to build a subject groups platform knowledge circle. Then，it constructed a subject groups interest model and calculated similarity between recommendation information and subject groups interest. Automatic and accurate recommendation of the scientific and technological information and the field monitoring express was realized on Subject Groups Integration Platform，which supported sustainable development of subject groups Integration platform.

〔**Key words**〕subject groups integration platform；knowledge circle；accurate information recommendation

近年来，随着物联网、云计算、社交网络等技术的快速发展，科技界也正在迅速建立传播、管理和处理全球知识的基础设施，构建将知识的交换、共享和处理作为所有应用和服务的核心的知识服务机制，其内涵是采用互联网技术和分布式的高性能计算环境来建立的一种全新的科学研究知识环境。康奈尔大学构建的交互式知识网络 Scholars @ Cornell 以可视化的形式揭示科研团队、领域专家的研究兴趣、科研动态、国际合作等科研要素，并提供知识关联和知识发现等功能[1]。英国曼彻斯特大学牵头建设的虚拟科研环境 myExperiment 被设计用来支持分享实验设计和工作流程[2]。美国的科研社交服务网站 ResearchGate 旨在推动全球范围内的科学合作，用户可以联系同行，了解研究动态，分享科研方法以及寻找工作机会[3]。

中国科学院在“十二五”期间以院所协同的方式开展了研究所群组集成知识平台可持续服务能力建设项目，面向课题组或实验室建立嵌入科研过程的知识服务与利用环境。截至目前，已经在全院范围内建成服务了 530 多个平台，培育和提高研究所在平台建设方法、资源组织利用、

收稿日期：2018-03-30
基金项目：中国科学院文献情报能力建设专项“全院科研群组平台运行及服务中心”（项目编号：Y6ZG7）；“面向科研群组的精准信息推荐”（项目编号：Y6ZG46）。
作者简介：王峰（1979-），男，副研究员，研究方向：信息平台建设。林丽珊（1962-），女，高级工程师，研究方向：文献情报和专利分析。刘毅（1972-），男，副研究馆员，研究方向：数字图书馆技术。

组织管理和人力等方面的能力，形成了院所协同的可持续服务机制，起到了良好的示范作用[4]。

但是，全球数据量出现爆炸式增长，数据与信息作为新兴战略资源。全球进入到一个以数据驱动社会创新的大数据时代，科学研究也正在向“第四范式——数据密集型科学发现”转变[5]。在大数据时代，科研人员想快速从海量数据信息中精准获取满足自身需求的信息资源却变得更加困难，信息过载问题日益突出，如何有效缓解此类问题，向科研人员精准推荐实用有效的科技信息已经成为一个重要课题。

本文在前期项目建设基础上，构建群组平台知识圈，开展面向科研群组的科技信息精准推荐服务，使得科研人员更方便快捷的使用文献情报服务，了解所需的科研信息、把握学科发展态势、查找科研资源、进行科研协作、开展科研实践、科研交流和协同等。

1 整体框架

科研群组是指具有共同或相似的研究领域或研究方向的科研人员集体。中国科学院由分布于全国各地的 100 多个研究所组成，各个研究所都有许多个不同研究领域的科研群组以相对独立自主的形式开展科研活动。所级群组平台建成后，存在着部署分散、相互之间缺乏有机联系、资源缺乏共享、未形成统一整体的知识服务网络等问题。在科技创新的信息需求不断扩展的大数据环境下，迫切需要建设一个群组平台知识生态系统，统一组织分散在各个所级群组平台的科研动态信息，集成相关领域的知识资源，汇聚形成一个知识不断流动的知识圈。在此基础上，向用户推荐最感兴趣的科技信息，打造一个真正建立嵌入课题组或实验室科研活动过程的知识环境，满足科研人员个性化、知识化的服务需求。系统整体框架如图 1 所示:

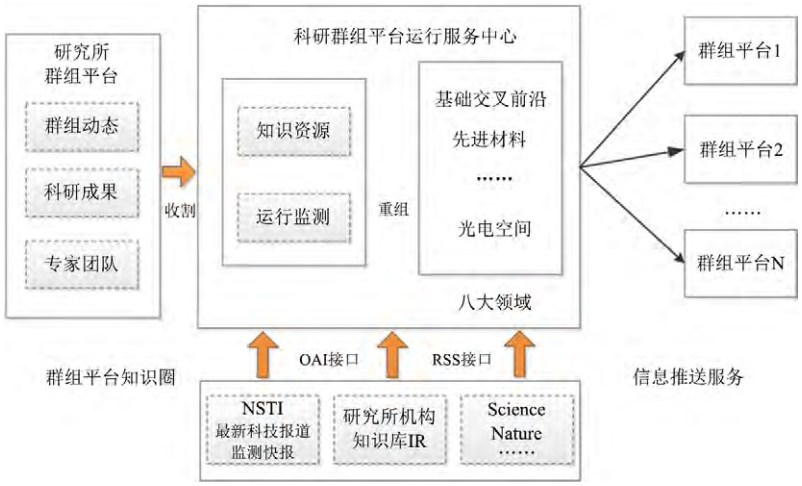

图 1 整体框架

1) 围绕《中科院“十三五”发展规划纲要》提出的“基础前沿交叉”、“先进材料”、“能源”、“生命与健康”、“海洋”、“资源生态环境”、“信息”、“光电空间”等八大创新领域，对建成的所级群组平台梳理分类，建立以科研群组平台运行服务中心为核心节点的群组知识圈。一方面定期收割研究所群组平台相关的综合科技信息，按领域范畴等统一组织；另一方面依托科技监测平台、机构知识库等第三方开放资源系统集成科技资讯、基金项目信息、专家信息、科研论文、专利文献等资源。2) 将获取到的知识资源按学科领域重组，在科研群组平台运行服务中心上开发信息推送服务，以数据接口的形式向所级群组平台自动投送，实现科技信息精准推送服务，促进知识共享流通。

2 以科研群组平台运行服务中心为核心构建群组平台知识圈

群组平台知识圈以科研群组平台运行服务中心为核心节点组织建设，突出科研群组平台面对一线科研人员的知识服务作用，汇聚包括科研过程中产出的论文、报告、论著在内的多方数据资源，实现知识资源的个性化定制利用，突出建设个性化的深度知识资源组织服务能力。

笔者分析评估了已经建成的 500 多个所级科研群组平台的数据资源情况，按照“基础前沿交叉”、“先进材料”等八大创新领域进行拆分和资源的二次加工，抽取科技资讯、科研项目、论文、专利等开放资源，集成科技监测系统、机构知识库等第三方资源，收割所级群组平台发布的群组动态信息，汇聚形成知识圈。

在群组平台建设中广泛使用的中国科学院集成信息平台 CASIIP 的原有功能基础上，遵循 JSR168 规范，保留整体系统框架，重新设计系统的页面布局、功能菜单和呈现样式，丰富知识圈功能，具体工作包括: 1) 利用已经部署在研究所科研群组平台的 RSS 数据接口，每天定时收割所级群组平台上发布的科研群组的研究动态、招生招聘、论

文专利等多种信息，在群组平台运行服务中心上集中展示。2) 通过 OAI 接口集成国家科技图书文献中心 NSTL 发布的编译报道和监测快报等科技信息，按照八大创新领域重新组织集成揭示。3) 根据不同的用户层级，设定用户浏览、获取、下载知识资源的权限控制方式。4) 提供实时运行监测服务，定期抽取各个科研群组平台的访问量、数据量等数据指标，引入可视化分析、预测分析等处理分析方法全面客观展现群组平台建设和服务的能力与效果。科研群组平台运行中心目前已经集成了群组动态2 417条、科技资讯23 693 条、科研项目信息 20 157 条、论文题录信息 190 069条、专家信息4 413条，首页服务界面如图 2 所示:

图 2 科研群组平台运行服务中心首页服务界面

3 面向科研群组的精准信息推荐

传统的推荐系统主要服务于个体，大量出现并成功应用于电子商务、位置服务、娱乐旅游等领域，但是面向团队、群组的推荐系统还没有形成统一的应用模式。面向科研群组的信息推荐不同于普通的个体推荐，科研人员由于研究背景，研究方向可能相同或相似也有可能存在一定差异，为此既要获取多个用户的兴趣偏好，还要对兴趣的差异性进行协调统一，以便推荐结果尽可能使同一个科研群组中尽可能多的成员满意。笔者提出了一种基于群组兴趣模型的信息推荐方法，在群组平台运行服务中心将科技资讯、领域监测快报发布为 RSS 服务，所级群组平台订阅上述服务并抽取具体信息的关键词向量，与表征科研群组研究兴趣的特征词向量进行相似度计算，根据计算结果对推荐信息过滤排序，从而实现信息的精准推送，信息推荐流程如图 3 所示。

3.1 构建群组兴趣模型

笔者提出了一种在综合分析科研群组的学科领域特征、发文情况、用户习惯以及所关注的信息源基础上，构建科研群组的用户兴趣模型的方法。中国科学院广州能源研究所（以下简称广州能源所）是国内从事清洁能源工程科学领域的高技术研究的重要科研单位，在前期项目中建成了多个科研群组平台。本文以该所为例详细说明群组兴趣模型的构建流程，如图 4 所示。

1) 广泛调研科研群组的需求、建议与意见，了解科研人员最希望获取的信息，形成共性需求。访谈课题组 PI 等重点用户，了解个性化需求。

2) 从中国期刊网 CNKI 下载最近五年研究所科研人员作为合作作者发表的论文，以 Excel 的形式保存题录信息，重点是抽取论文的关键词。

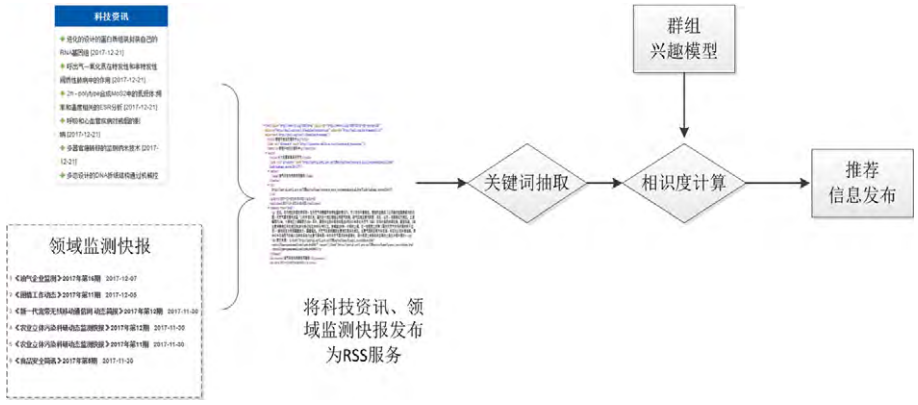

图 3　信息推荐流程图

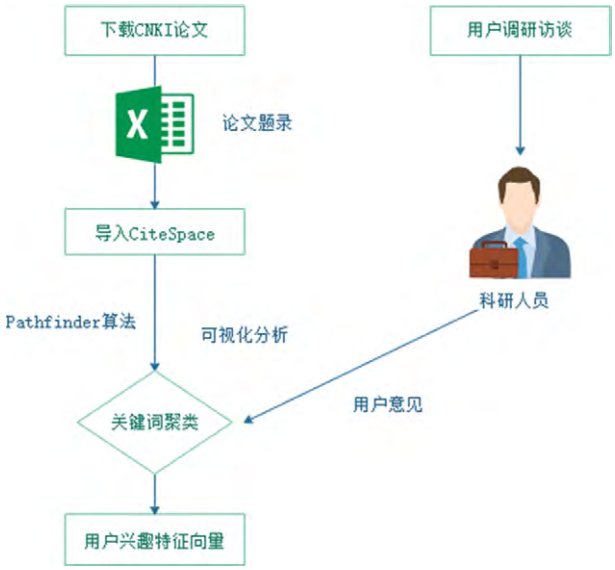

图 4　用户兴趣模型构建

3）**应用可视化文献分析软件** CiteSpace **的** Pathfinder **算法对论文关键词进行聚类分析。**CiteSpace **能够显示一个学科或知识领域在一定时期发展的趋势与动向，形成若干研究前沿领域的演进历程**[6]**。**Pathfinder **算法是一种灵活有效且高效的网络简化算法，在知识图谱发现中有较多应用，是** CiteSpace **中对于共现矩阵的可视化使用的基本算法。关键词聚类的可视化分析结果如图** 5 **所示。**

4）**将论文关键词聚类结果提交给专家人工判读，并与调查、访谈的结论反复比较，抽取共性因素，去除关注度下降的关键词，最终以特征向量的形式来表现科研群组的用户兴趣。**

分析结果如下：

$\gamma_{\text{用户兴趣特征向量}}$={“水合物”，“热解”，“生物”，“柴油”，“天然气”，“合成气”，“甲烷”，“太阳能”，“燃料”，“电池”}

图 5　论文关键词聚类结果

3.2 关键词抽取

关键词抽取是指从文本中自动抽取出若干有意义的词语或词组，用以反映文本的主要内容。关键词抽取技术主要分为 3 类：基于统计特征的算法、基于主题模型的算法和基于词图模型的算法。

经过调研和分析，选择 TextRank 算法来抽取向所级群组平台推荐信息的关键词向量。TextRank 算法是一种典型的基于词图模型的算法，其提出受到 Google 的 PageRank 算法启发。TextRank 算法针对文本里的句子设计权重，利用局部词汇之间关系对后续关键词进行排序，直接从文本本身抽取，不需要先期训练过程，实现相对简单[7]。

本文以标题为“生物柴油生产的微反应器技术”的一段科技资讯为例来说明关键词向量抽取的效果。

资讯原文 “由于全球能源需求的增加和温室气体的增加，生物柴油作为一种替代燃料已被人们所接受，因为生物柴油的生物可降解性、低环境有害影响、更好的废气排放质量和可再生能力。生物柴油是一种由长链脂肪酸组成的单烷基酯，也被称为脂肪酸甲酯。酯交换是生产过程中最常用的技术。但传统的生物柴油技术也有其不足之处。过程强化技术可以克服这些缺点。一些新型的反应器，如微通道反应器、静态混合器、振荡流反应器和旋转管反应器已经被开发出来，以改善质量转移和混合。由于高表面积/体积比和短扩散距离的影响，这些技术可以实现快速和高的反应速率，从而强化了酯交换过程。酒精对甘油三酸酯的摩尔比、微通道大小、停留时间、反应温度、搅拌机理、催化剂等因素影响了生产过程。尽管生物柴油的生产已经在几个国家被商业化，但它仍然需要一个清洁、有效和环境友好的技术，以使其具有成本效益并提高其对传统矿物燃料的能力。然而，微反应器技术已经被证明是达到这一目的的基准。本文综述了生物柴油生产中使用的不同类型的微反应器，以及影响微反应器中生物柴油生产的参数。本文讨论的微反应器技术旨在通过减少反应时间到分钟来改善生产过程。”

抽取的关键词向量：

$\gamma_{\text{关键词向量}}$={“生物柴油”，“燃料”，“微反应器”，“过程强化技术”，“酯交换”}

3.3 相似度计算

本文通过计算群组兴趣特征向量和推荐信息关键词向量之间的相似度，来表征面向科研群组推荐的科技信息与科研人员研究兴趣的准确度。因为科技资讯属于文本信息，所以借鉴了文本相似度的计算方法。常见的文本相似度计算方法分为基于统计或者语料库的方法和基于世界知识的方法。基于字符串的方法也称作“字面相似度方法”，其中较为典型的方法包括最长公共子串，编辑距离等。基于世界知识的方法是指利用具有规范组织体系的知识库计算文本相似度[8]。笔者结合上述文本相识度计算方法的优点和特点，引入同义词、近义词、上位词、下位词的概念比较两个向量编辑距离得到语义相似度。编辑距离，又称 Levenshtein 距离，是指两个字符串之间，由一个转成另一个所需的最少编辑操作次数，如果它们的距离越大，说明它们越是不同。引入同义词、近义词的概念后，如果词语之间存在同义、近义、上位、下位的关系，两者之间的距离则由语义关系权重决定，例如“石油”和“原油”为同义词，两者之间的距离定义为；“石油”和“汽油”为上下位词，两者之间的距离定义为 0.9。

选取 20 篇来自 NSTL 重点领域信息门户的科技资讯与本文 3.1 节计算出的广州能源所科研群组用户兴趣特征向量进行语义相似度计算，同时采用专家咨询的方式获得人工对于准确度的判断。相似度用 [0,1] 之间的实数表示，0 表示两个概念完全不同，1 表示两个概念语义相同，计算结果如表 1 所示。经专家判读，部分相似度低于 0.25 的信息与广州能源所的研究方向差异较大，可以直接忽略，相似度大于 0.5 的信息值得重点推荐。

表 1 相似度计算结果

标 题	关 键 词	相似度
强大的能量流流经麻省理工学院	能源，能量流，MIT，技术，机会	0.1420
核反应堆退役是一个长期且代价高昂的过程	核反应堆，退役，过程，方法，设备	0.2729
Covanta：都柏林 EFW 设施已经开始商业运作	Covanta，工厂，设施，都柏林，运营	0.1029
国际能源机构国家能源政策——2017 年希腊审查	能源，希腊，政策，可持续，再生	0.2739
美国能源部安排生物量咨询委员会会议	美国，能源部，生物，物质，公报	0.5094
探索频道的纪录片聚焦生物柴油	生物柴油，能源，美国，油脂，企业家	0.7584
3 个问题：Robert Granetz 在核聚变研究上	等离子体，粒子，核聚变，反应堆，电子	0.2302
研究生获得能源计算奖学金	奖学金，计算，工程，物理，研究生	0.3321
中国和越南在核安全问题上的合作	越南，中国，核安全，备忘录，合作	0.2302
让可再生能源更可行	可再生能源，电池，储存，风能，太阳能	0.6064

表 1（续）

标　　题	关　键　词	相似度
从生物质和风能的生产中产生的柴油—通过使用费施托—托普施的过程	生物质，风能，柴油，燃料，储存	0. 5678
通过甲基环己烷脱氢制氢催化膜反应器的模拟和设计	反应器，催化，模拟，甲基环己烷，氢	0. 5282
熔融氢化 nb-si 合金的微结构和机械性能	氢，压缩，高温，合金，nb-si	0. 1288
德国能源公司的辉煌成就，以及接下来需要做的事情	德国，能源，政府，转型，再生	0. 2224
研究了风速对二氧化碳在生物吸收过程中二氧化碳损失的影响	二氧化碳，风速，大气，生物，反应器	0. 5121
佩里的行动是为了防止掠夺性定价，把核和煤炭竞争推到市场之外。价值弹性和管道独立性的价值	核，煤炭，能源，规则，市场	0. 5121
能源机构—工作组解除拆除和拆除工作（WPDD）	能源，政策，WPDD，组织，专家，监管	0. 2211
用模拟余热加热管式反应器中甲醇蒸汽重整的流动和操作参数的影响。	甲醇，蒸汽，摩尔，参数，反应器	0. 5076
基社盟在推进藻类生物燃料方面的项目	藻类，项目，生物，燃料，能源	0. 7506
ABB 展示了电动汽车充电解决方案的范围	充电，汽车，ABB，解决方案，网络	0. 2978

3. 4　应用效果

本文结合项目研究需求，首先在广州能源所知识服务集成平台中实际应用。针对广州能源所的一三五重大突破方向“新能源与可再生能源领域的研究与开发利用”，“生物质能源高值化转化与规模化利用”，如图 6 和图 7 所示，重点推送了可再生能源、生物能源等领域的科技资讯和《可再生能源领域监测快报》。从 2018 年 1 月 24 日到 2018 年 3 月 21 日自动推荐能源科技信息 321 条，系统自动过滤了相似度较低的 38 条信息，实际推送了 283 条信息，重点推荐与研究所研究方向相似的 40 条信息和《可再生能源领域监测快报》12 篇。

图 6　科技资讯推荐服务界面

4　结束语

实践应用表明，本文通过精确的数据分类、组织构建群组平台知识圈，分析科研团队的用户研究方向与兴趣，形成用户兴趣模型，实现向科研群组推送与其研究方向相关的知识资源。从而提高资源和数据的应用价值，提升用户浏览兴趣，改善服务体验，增加用户粘度，支持科研群组平台的可持续发展。但同时，本文也存在一定不足，群组兴趣模型的构建过程相对复杂，随着科研活动的深入和新的创新领域不断涌现，科研人员的研究兴趣也会随之发生转移，以上问题需要后续深入研究，期望能够借助人工智能、机器学习的算法，完成群组兴趣模型的自动构建。

图 7 《可再生能源领域监测快报》推荐服务界面

参 考 文 献

[1] Scholars @ Cornell [EB/OL]. https://scholars.cornell.edu/，2018-03-15.

[2] ResearchGate [EB/OL]. https://www.researchgate.net，2018-03-15.

[3] myExperiment [EB/OL]. https://www.myexperiment.org/home，2018-03-15.

[4] 杨志萍. 群组集成知识平台可持续发展能力建设应用实践[J]. 现代图书情报技术，2012，(7/8)：32-37.

[5] 潘教峰，张晓林. 第四范式：数据密集型科学发现[M]. 北京：科学出版社，2012.

[6] CiteSpace：Visualizing Patterns and Trends in Scientific Literature [EB/OL]. 2017-12-26.

[7] Mihalcea R，Tarau P. TextRank：Bringing Order Into Texts [C]. Association for Computational Linguistics，2004.

[8] 陈二静，姜恩波. 文本相似度计算方法研究综述[J]. 数据分析与知识发现，2017，(6)：1-11.

(责任编辑：陈 媛)

(上接第 73 页)

[10] Kim H N，Saddik A E. Exploring Social Tagging for Personalized Community Recommendations [J]. User Modeling and User-Adapted Interaction，2013，23 (2-3)：249-285.

[11] 冯勇，李军平，徐红艳，等. 基于社会网络分析的协同推荐方法改进[J]. 计算机应用，2013，33 (3)：841-844.

[12] 孙甲申，王小捷. 一种用于社会化标签推荐的主题模型[J]. 北京邮电大学学报，2014，37 (3)：38-42.

[13] 易明，邓卫华，徐佳. 社会化标签系统中基于组合策略的个性化知识推荐研究[J]. 情报科学，2011，(7)：1093-1097.

[14] 刘健，尹春霞，原福永. 基于非结构化 P2P 网络用户模型的协同过滤推荐机制[J]. 山东大学学报：理学版，2011，46 (5)：28-33.

[15] 魏建良，琚春华. 基于社会化标注的用户协同模型研究[J]. 情报学报，2012，31 (3)：281-288.

[16] Liu K，Fang B，Zhang W. Exploring Social Relations for Personalized Tag Recommendation in Social Tagging Systems [J]. Ieice Transactions on Information & Systems，2011，94-D (3)：542-551.

[17] 李慧宗，周姣，王向前，等. 融合社会关系的用户标签主题模型[J]. 情报杂志，2017，36 (3)：165-172.

(责任编辑：孙国雷)