大数据环境下科技查新用户深层次精准服务推荐构想*

梅 梅 唐小利 张 玢 王 超

(中国医学科学院医学信息研究所/图书馆 北京 100005)

〔摘要〕 通过科技查新工作数据和互联网数据的结合，进行查新用户的潜在需求、兴趣动向、偏好习惯的精准分析，从大数据技术方面分析对于海量用户数据设计有针对性的深层次服务的可行性，提出基于科技查新用户开展深层次精准服务推荐的框架和方法，探讨为科技查新用户提供深层次精准服务的机遇和挑战。

〔关键词〕 科技查新；大数据；精准服务

〔中图分类号〕R-056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2016.06.017

Concept of Deep-level and Precise Service Recommendation for Users of Sci-tech Novelty Retrieval in the Environment of Big Data

MEI Mei, TANG Xiao-li, ZHANG Bin, WANG Chao, Institute of Medical Information/Library, Chinese Academy of Medical Sciences, Beijing 100005, China

〔**Abstract**〕 By integrating data of sci-tech novelty retrieaval and internet data, the paper makes a precise analysis of the potential demands, interest, interest, tendency, preferences and habits of users of novelty retrieval. From the perspective of big data technology, it analyzes the feasibility of designing targeted and deep-level services for mass user data, puts forward the framework and method of conducting deep-level and precise service recommendation based on users of sci-tech novelty retrieval, and studies the opportunity to provide these users with deep-level and precise services.

〔**Keywords**〕 Sci-tech novelty retrieval; Big datu; Precision services

1 引言

科技查新是国家为客观正确地判别科研成果的新颖性避免科研课题重复立项而设立的一项工作，由具有科技查新资质的查新机构承担完成。查新机构根据查新委托人提供的需要查证新颖性的科学技术内容，按照科技查新规范操作，提供科技查新服务[1]。早在 1990 年，数据仓库之父比尔·恩门(Bill Inmon) 就开始关注大数据，但最早提出“大数据”概念并被广泛传播则始于 2008 年 9 月，当时《自然》杂志以专刊的形式，详细介绍了大数据的起源、特征、组织形式、长久保存和利用等。IBM 认为大数据具有 4V 特性：数量、多样性、速度和真实性。其实，大数据描述的是随着数据量和数据类型激增而出现的一种大规模、多样化的数据集，及其对数据集高速采集、分析、处理以提取知识价值的技术架构与过程[2]。《科学》杂志推出过数据处理专刊，其中认为对科学学科而言，大数据

〔修回日期〕 2015-10-09

〔作者简介〕 梅梅，馆员；通讯作者：唐小利，研究馆员。

〔基金项目〕 国家科技图书文献中心委托专项任务（项目编号：2014XM052，2014XM056）。

既是挑战也是机遇[3]。虽然科技查新可以为科技成果的鉴定、评估、验收、转化、奖励等提供客观的依据，但从数据使用的角度来看，科技查新工作还只是停留在数据使用的最初阶段，如何让数据转变为信息，再让信息转变为智慧是数据应用面临的主要问题。大数据环境下需要从查新工作的工作数据、数据库中的数据及网络数据中获取更多信息，将获取到信息和查新工作人员的经验知识相结合，利用智能知识管理[4]手段发现这些信息知识中所蕴含的规律，从查新用户过往的查新委托、查新习惯模式和倾向中寻找启示，并加强用户需求分析与相关交互数据的利用，为查新用户提供深层次精准信息服务提供参考建议。对于科技查新部门而言，如果能进行大量内外部数据的搜集、整理、组织、分析和决策，就有可能获取到查新用户的精准服务需求，有针对性拓展深层次信息服务，而且这种经过二次加工集成获取的用户需求将更具有新颖性和实用性。大数据环境下实现科技查新用户深层次精准服务推荐，核心在于多维用户数据的融合、用户行为特征的提取、推荐模型构建、及数据驱动的科技查新服务策略。

2 大数据环境下开展精准服务推荐原则

2.1 需求匹配原则

基于科技查新用户提供的精准服务的前提是基于用户的行为数据、查新工作数据等并运用相关数据挖掘方法和技术手段对用户和服务进行精准的匹配。科技查新工作具有较强的政策性，用户查新多数都是由于政策导向基于科研立项或成果评奖的要求需要查新，因而在单一使用科技查新工作数据进行用户精准服务推荐时必然受到限制以至于匹配不准确，但大数据环境下，由于大量的互联网数据的产生，而且所有人在互联网上的活动都是可以留下痕迹的，因此通过科技查新的工作数据、数据库的文献专利等数据及互联网数据的融合，根据学科背景、所在单位等将用户进行群体划分后，然后通过捕捉相关群体用户的互联网数据，通过一定的算法和技术，依据他们的行为特征进行匹配服务推荐。

2.2 精准性原则

鉴于国家政策和外部环境的不确定性、科技查新工作人员工作能力强弱的不确定性、用户需求变化的不确定性、用户管理的随机性、数据来源的不稳定性等，存在用户群信息不确定性和个人用户信息的不确定性，要实现精准服务的一个最大问题就是服务推荐过程中的不确定性因素。而又由于不确定性是客观事物存在或发展过程中的一种客观属性。客观世界当中的大多数现象都具有不确定性[5]，因此要综合运用关联规则、聚类、协同过滤、图挖掘、组群挖掘、概念相似度、动态描述逻辑和序列挖掘等方法尽可能掌握服务对象的需求，进行精准服务推荐。

2.3 主动服务原则

大数据的核心是建立在相关关系分析法基础上的预测，这种预测要解决的问题是“是什么”，而不是“为什么”，这种预测是在大数据及其创新的分析技术支撑下的非因果关系的预测[6]。大数据环境下基于科技查新用户精准服务推荐需要通过相关关系分析所有获取的数据，根据分析结果改变原有的查新工作被动服务原则，为用户在新的视角下提供主动服务。这种主动服务必须基于大数据的环境，通过各种数据关联和算法技术，预测科技查新用户的深层次服务，给用户提供相关精准服务推荐，以满足用户的需求，而非传统模式下靠用户拉动的服务理念。

3 深层次精准服务推荐构想

大数据环境下基于科技查新用户提供深层次精准服务推荐先从整体进行用户群组的服务推荐，然后再针对不同群组中的用户进行个性化服务推荐。由于科技查新用户存在很强的异质性，不同类型用户对查新服务的要求和标准也存在显著差异，因此有必要针对不同层次的用户，为其推荐更加精准的产品和服务。利用大数据相关技术可以全方位了解科技查新用户的观点评论、需求偏好和行为习惯，通过相关数据挖掘方法，帮助查新机构构建深层次

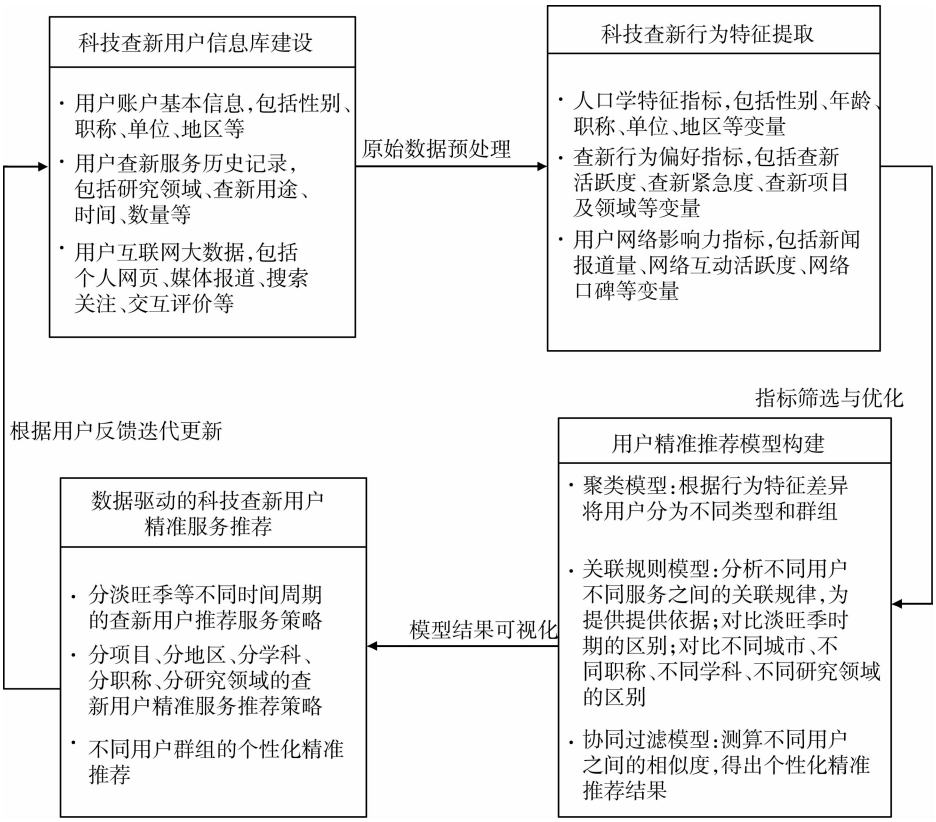

图1 大数据环境下基于科技查新用户开展深层次精准服务推荐框架

精准服务推荐体系。大数据环境下的深层精准服务推荐体系建立需要4 个主要步骤，分别是科技查新用户信息库构建、查新行为特征提取、用户精准推荐模型开发、数据驱动的科技查新服务策略。

3.1 科技查新用户信息库建设

用户信息库的建设是精准服务的基础工作，信息库应该既包括查新机构内部数据，也包括外部互联网大数据，不同来源的数据具有不同的含义，主要包括如下几类：（1）科技查新用户的账户数据，包括人口学特征，如姓名、单位、职称、所在城市等。（2）科技查新用户的服务历史数据，包括该用户过去提交查新请求的时间、数量、研究领域、查新用途、是否加急、是否有修改反馈等。（3）科技查新用户互联网大数据，包括查新用户的个人网页数据、新闻媒体报道数据、微博数据、医学垂直网站交互及评价数据、搜索引擎数据等。

3.2 科技查新用户行为特征指标提取

（1）人口学特征。性别、年龄、职称、城市、单位类型，除了年龄是数值型变量外，其他都是分类变量。（2）查新行为。研究领域、查新用途、查新所需数据库、查新总次数、最近一次查新截至当前的时间间隔，其中两个编码是分类变量，其他是数值型变量，查新总次数反映了该用户查新活跃度，最近查新时间间隔可以一定程度反映申请人项目申请的连续性。（3）用户网络影响力。新闻媒体报道量、网络互动活跃度、网络口碑，前两个都是数值型变量，最后一个是分类变量。媒体报道量反映该用户的官方提及度或名望，网络活动活跃度可以用该用户网络发帖回帖量来表征，反映其交流意愿；网络口碑通过网络对该用户评价的文本分析而得出，分为正向、负向、中性3 个类型。

3.3 用户精准服务推荐模型

在上述行为特征指标基础上，首先通过聚类模型把所有查新用户样本划分为不同类型的群组，将各指标进行标准化处理，采用 Kmeans 等聚类算法进行聚类分析，调整合适的参数得到聚类结果。然

后针对每个用户群组及不同用户群体中的用户，分别应用关联规则模型和协同过滤模型展开精准服务推荐，两个推荐模型的原理和用途如下：

3.3.1 关联规则模型 是一种发现顾客多种需求，并通过满足其需求而推荐多种相关服务或产品的方法。基于关联规则的个性化推荐系统具有能够发现顾客新的兴趣点和不需要领域知识的优点[7]。基本原理是，假设 I = {i1, i2, …, im} 是 m 个不同项目的一个集合，称之为项目集，在查新关联分析中，可以将用户的查新服务、用户试用过的其他服务、用户行为特征指标作为 i1, i2, …, im 代入项目集，X 和 Y 是项目子集，定义支持度表示在项目集中同时出现 X 和 Y 的比率，置信度为在出现 X 的情况下又出现 Y 的比率。通过对各个客户群的共性特征进行挖掘，得到客户群共性特征的关联规则，将客户群数值特征矩阵和客户群共性特征的关联规则存储起来，形成客户群共性特征库[8]。在此基础上，分析用户个体行为与用户群体的共性特征差异，判断差异值对服务的影响，确定单个用户的影响因素，并基于这种差异性实现对单个用户的服务推荐。通过分别计算不同服务之间、行为特征与服务之间的支持度和置信度结果，可以分析两个不同查新服务的关联规律，为提供其他服务寻求依据，还可以得出不同时期（淡旺季）查新服务的关联差异，不同职称、不同学科、不同研究领域对于不同类型信息服务的需求。

3.3.2 协同过滤模型 是一种基于用户相似度计算的个性化服务推荐方法。基本原理是，基于邻居用户的兴趣爱好预测目标用户的兴趣偏好。算法首先采用统计技术寻找与目标用户有相同喜好的邻居，然后根据目标用户邻居的喜好产生向目标用户的推荐[9]，具体步骤如下：（1）表示：输入数据通常可以用一个 m × n 的矩阵 Rm × n 来表示，这个矩阵也称用户 – 项目矩阵。Rm × n 表示 m 个用户对 n 个商品的评价，ri, j 表示用户 i 对商品 j 的评价值[10]，见图 2。（2）邻居的形成：其核心是为一个需要推荐服务的目标用户寻找最相似的“最近邻居”集（Nearest – neighbor），即：对一个用户 u，要产生一个根据相似度大小排列的“邻居”集合 N = {N1, N2, …, Nk}，从 N1 到 Nk，相似度 sim（u, N1）> sim（u, N2）> … > sim（u, Nk）。通过计算目标用户 1 和其他用户之间的相似性（比如计算欧几里德距离），以点 1 为中心的 2 – 6 个最近用户被选择为邻居[11]，见图 3。（3）产生推荐：目标用户的“最近邻居”集产生后，可计算两类结果：用户对任意项兴趣度的预测值和 Top – N 形式的推荐集。在查新用户的协同过滤模型中，可以将用户购买的产品或服务、用户行为特征指标赋值给 Item 向量，用来作为计算用户相似度的特征变量，通过模型运算即可以得出个性化的精准推荐结果。因此，基于关联规则模型的结果，结合查新机构的人员和资源状况，可以得出不同时间周期（淡季旺季）的查新用户精准服务推荐策略，结合协同过滤模型的结果，还能通过数据运算结果得出不同查新用途申报用户、不同地区用户、不同职称用户、不同学科领域用户及用户群体的个性化服务推荐列表。查新机构可以选择一部分查新用户对上述推荐策略进行测试，及时记录用户反馈，并将其补充到用户信息库中用来迭代改进，最终使得推荐模型越来越精准，待达到理想准确率之后再向所有用户应用。

用户（User）	项目（Item）				
	i_1	i_2	i_3	i_4	i_5
u_1	2	3		4	
u_2	1		2	1	1
u_3	4	4	5	4	3
u_4	1		3		
u_5		2		5	

图 2 用户 – 项目矩阵

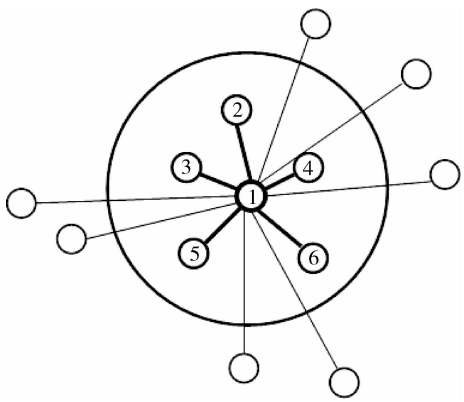

图 3 邻居的形成

4 结语

大数据环境下，科技查新部门也应该具有市场意识[12]，这也是未来科技查新部门面对的发展和挑战。通过从科技查新工作数据中找到切入点，充分利用查新工作的业务数据并结合互联网数据，使用数据挖掘算法和相关大数据技术，能按照用户的真正需求尤其是潜在信息需求，为不同用户提供相应的深层次服务推荐，并通过与用户建立的微博、微信等沟通渠道在后续服务推荐中全程与用户保持互动，根据用户的需求不断调整相应的服务推荐列表，不断完善信息服务体系，让每一个用户群组及个人用户充分享受深层次的服务推荐。

大数据环境下也可以完全贯彻“以用户为中心”的思想，通过现有的查新工作数据了解现有的服务过程发生了什么，利用数据对科技查新工作与用户的交互关系进行数据挖掘、分析和预测可能发生的信息行为。加强用户研究与交互互联网数据的利用，对用户数据进行深度分析并建立用户模型，开展精准服务、知识关联服务，提供预测性信息服务产品[10]。从而逐渐转变查新部门的被动服务模式，将查新部门打造成为学科服务中的重要关键部门，为学科服务提供有利而关键的重量级支持。

参考文献

1 李俏. 基于大数据环境下的科技查新服务研究［J］. 决策与支持，2015，(7)：81.

2 刘高勇，汪会玲，吴金红. 大数据时代的竞争情报发展动向探析［J］. 图书情报知识，2013，(2)：105－111.

3 GINSBE R G J. Detecting Influenza Epidemicsusing Search Engine Query Data［J］. Science，2009，(457)：1－5.

4 李兴森，石勇，张玲玲. 从信息爆炸到智能知识管理［M］. 北京：科学出版社，2010.

5 李德毅，刘常昱，杜鹢. 不确定性人工智能［J］. 软件学报，2004，15（11）：1583－1594.

6 黄红梅. 大数据时代学科服务理念创新［J］. 情报资料工作，2015，(3)：68－70.

7 Li J，Xu Y，Wang Y F，et a1. Strongest Association Rules Mining for Efficient Applications［C］. Proceeding of the Fourth IEEE Conference on Service Systems and Service Management，2007：502－507.

8 皮佳明. 基于用户兴趣变化的协同过滤推荐算法研究［D］. 昆明：云南财经大学，2014.

9 李冰. 基于二次匹配的精准服务推荐研究［D］. 武汉：武汉理工大学，2014.

10 吴婷. 协同过滤技术在电子商务推荐系统中的应用和研究［D］. 武汉：武汉理工大学，2009.

11 周强. 基于用户的协同过滤推荐算法研究［J］. 南昌高专学报，2006，(6)：88－90.

12 吴慧敏. 大数据与图书馆信息服务新构想［J］. 图书馆与理论，2015，(2)：14－17.