

面向NSTL的个性化推荐系统构建研究*

□ 胡小丰 赵瑞雪 鲜国建 寇远涛 / 中国农业科学院农业信息研究所 北京 100081

摘要: 随着数字图书馆文献资源数量的高速增长,提升个性化信息服务的质量与效率,逐渐成为当前数字图书馆领域研究人员重点关注的方向。个性化推荐作为其中用户接受度较高的一种方式,呈现了快速发展的态势。文章介绍了当前国内外文献服务平台以及电子商务平台个性化推荐的应用情况,并结合NSTL网络服务系统的个性化信息服务现状,从体系结构设计、系统功能升级、核心模块设计与实现方法三方面对NSTL网络服务系统进行了个性化推荐系统设计。

关键词: 信息服务, 个性化服务, 推荐系统, NSTL

DOI: 10.3772/j.issn.1673—2286.2012.12.009

1 引言

进入21世纪以来,互联网技术发展迅速,创新层出不穷,网络资源也随之呈现爆炸式增长的态势,如何从海量的多元数据和信息中,快速、高效地获取有用的信息,如何从迅速膨胀的信息资源中获取最新的信息,成为了人们关注的重点领域。信息获取主要来自用户自助检索与系统推荐等方面,如何智能地呈现检索结果、高效地进行个性化推荐,成为能否满足用户信息需求的关键方面,也是衡量信息服务是否细致、精准、个性化的重要标准。

在个性化信息服务领域,有两大类服务平台,一类是文献数据库平台,国内有中国知网、万方数据等平台,国外有医学数据库平台PubMed等;另一类是电子商务平台(主要为B2C),国内目前重要的B2C电商平台有当当、京东、淘宝天猫等,国外则有Amazon这一全球性的电子商务平台。以中国知网、万

方数据为主的文献数据库平台在个性化推荐方面作出了探索和实践,围绕“以用户为核心”这一理念,挖掘用户行为信息,取得了很好的应用效果。电子商务平台以Amazon为代表,依托海量级用户数据以及云计算技术,针对单个用户、用户群的推荐形式丰富,能够覆盖与用户相关的各个方位,且更新及时,这也为其带来了巨大的商业价值。据统计,在基于电子商务的销售行业使用个性化推荐系统后,企业销售额平均提高2%~8%^[1]。

国家科技数字图书馆(NSTL)作为国内权威科技文献信息资源收藏和服务中心,面向全国开展科技文献信息服务。NSTL服务项目主要包括文献检索、期刊浏览、全文文献、引文检索、代查代借、参考咨询、自助中心、预印本服务等。作为国家级现代信息技术应用示范区,NSTL在发展历程中不断引入现代化的信息技术,为更好地服务广大用户而不断创新。本文针对NSTL

目前已有的个性化推荐功能,结合最新的个性化推荐技术,提出了个性化推荐系统升级的设计方案。其中,可以引入个性化推荐服务功能的模块有文献检索、原文传递、代查代借、自助中心。

2 个性化推荐的应用现状

随着文献服务平台以及电子商务平台的快速发展,现代信息技术已逐渐大范围、深度应用在提升信息服务质量、提高用户满意度等方面。其中个性化推荐技术在上述两种平台中得到广泛的应用。

2.1 文献服务类

万方数据、中国知网作为国内优秀的文献数据库平台,他们的个性化推荐服务开展较早,并成为引导读者用户的重要功能。万方数据的个性化推荐内容丰富,主要分为

* 本文系国家“十二五”科技支撑计划项目课题“信息资源自动处理、智能检索与STKOS应用服务集成”(编号:2011BAH10B05)研究成果之一。

三类：学者推荐、检索词及词条推荐、相似文献推荐。中国知网的个性化推荐服务与万方数据有很多相似的功能，包括相关作者推荐、相关研究机构推荐、相似文献推荐。同时还提供了同行关注文献推荐，这是因为同行关注较多的一批文献具有科学研究上的较强关联性。相同导师论文推荐功能则向用户推荐与该论文作者相同导师的其他学生所发表的硕士或者博士论文^[11]。PubMed作为国际知名的开放获取医学数据库，同样重视个性化推荐服务的应用。在用户输入检索词后，命中文献列表页面向用户推荐相关检索词、题目含有该检索词的文献、最近检索记录。同时PubMed也推荐相关引文，该功能采用了词语权重算法，按照相关程度从高到低对相关引文进行排序^[10]。

以上三种平台推荐功能如表1所示。

在文献数据库平台中，由于文献资源与电子商务网站中的商品资源属性不同，推荐相比电子商务网站有自己的特点，总结来看主要有基于文献资源进行推荐，以及基于用户群的推荐。例如，在万方数据、中国知网中都出现了基于用户群即读者群的推荐：本文读者也浏览过。同行关注和文献推荐两种场景推荐原理相同，即与目标文献同时被多数读者（用户）关注的文献推荐，该推荐场景分析了用户的浏览行为并进行用户群划分，然后作出推荐。这也符合个性化信息服务的发展方向，逐渐走向以用户为核心，深入挖掘用户个人信息以及行为信息，进行更有针对性、更符合用户兴趣的推荐。

由于与文献服务类平台中的资源相似，都属文献资源，所以当前

文献服务类平台中的个性化推荐功能对NSTL的个性化推荐服务升级有重要的借鉴意义。

2.2 电子商务类

在电子商务网站中，个性化推荐功能比较丰富，通过对Amazon、当当网和京东等网络平台的调研分析，总结出其个性化推荐主要分为两大类：基于用户群行为的推荐以及基于单个用户行为的推荐。用户群主要分为两种：浏览同一商品的用户群、购买同一商品的用户群，针对两种不同的用户群进行推荐。对非登录用户，在首页会实时根据该用户浏览的商品记录进行实时推荐，该功能需要大量的实时计算，主要应用在Amazon中。对登录用户，进入用户管理界面后，会根据该用户的购买记录、评分进行综合推荐。其中在Amazon中，用户可以参与到改善推荐的操作中，通过对购买过的商品进行打分或重新打分、勾选“不用于推荐”来影响推荐结果^[2]。电子商务平台的推荐功能总结如表2。

3 NSTL个性化推荐体系结构

3.1 总体框架

面向NSTL的个性化推荐在NSTL原有推荐功能基础上，将个性化推荐技术引入文献服务系统中，进一步扩充其推荐功能，旨在提高其个性化信息服务水平，完善系统的服务功能，向广大用户提供个性化、专业化、优质的信息服务。其总体框架如图1所示，包括数据支撑层、核心业务层、服务层。

数据支撑层：NSTL数据支持，以接口或数据库授权方式可以访问到，为个性化系统提供基本的数据支持。

核心业务层：系统中重要的层次，实现个性化服务的核心业务。主要包括个性化数据仓储和五个模块（重排序模块、同行行为分析模块、兴趣推荐模块、文献相似模块、元数据共现模块）组成。

服务层：主要为NSTL系统用户提供的个性化相关服务，包括六个场景：检索结果重排、共同浏览推荐、共同订购推荐、用户兴趣推荐、相似文献推荐、共现作者及作者机构推荐。

3.2 推荐模型

（1）整体模型设计。对本系

表1 文献服务平台推荐功能

	万方数据	中国知网	PubMed
相关作者推荐	√	√	
相关研究机构推荐		√	
相似文献推荐	√	√	√
相关检索词推荐	√	√	√
相关词条推荐	√		
相关博文推荐	√		
本文读者也浏览过	√	√	

表2 电子商务平台推荐功能

	推荐功能	卓越网	当当网	京东网
基于用户群行为的推荐	浏览了同一商品的用户群最终购买的商品	✓	✓	✓
	浏览了同一商品的用户群还浏览了哪些其他商品		✓	✓
	购买了同一商品的用户群还购买了哪些其他商品		✓	✓
	购买了同一商品的用户群还浏览了哪些其他商品	✓	✓	✓
基于单个用户行为的推荐	首页实时推荐	✓		
	用户管理中心推荐	✓	✓	✓
改善推荐		✓		
邮件推送		✓	✓	✓

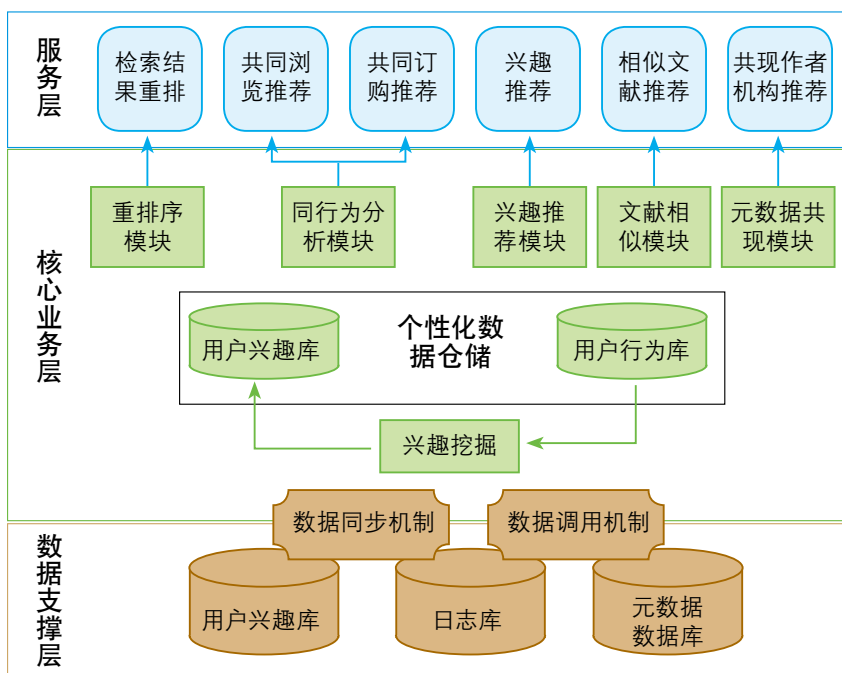


图1 系统总体框架

统模型设计由四部分构成,包括个性化推荐模型、用户兴趣模型、用户行为日志模型、NSTL文献资源模型。个性化推荐模型的输入为用户兴趣模型以及NSTL文献资源模型,输出为推荐的文献、作者、机构。用户兴趣模型的输入为用户行为日志模型以及NSTL文献资源模型,经计算后输出为兴趣关键词、权重、来源、更新时间等。NSTL文献资源模型将NSTL文献资源结构

化表达,以便与用户兴趣进行匹配完成推荐。

用户行为日志模型输入为用户显式行为和隐式行为,从NSTL文献资源模型获取相应的文献资源,为用户兴趣模型构建提供数据依据^[3]。

(2) 个性化推荐模型。个性化推荐模型输出了以不同形式推荐的资源,包括检索结果重排序、相似文献、相似机构、相似作者、用户感兴趣的文献、订购此文献的用户还订购过、浏

览此文献的用户还浏览过。

(3) 用户兴趣模型。用户兴趣模型由六元组构成^[4]: {兴趣关键词, 兴趣来源, 兴趣频次, 兴趣创建时间, 兴趣更新时间, 兴趣权重, 如图2所示。

该模型利用用户的注册信息进行初始化,根据用户几种行为日志进行定期、增量更新,主要利用文献ID、日志类型、操作时间,结合NSTL资源模型等;注册用户可对兴趣自主管理,包含添加、删除。

(4) NSTL文献资源模型。NSTL文献资源模型由一组元素集合构成: {ID号, 标题, 摘要, 关键字, 作者, 机构, 资源类型, 分类号, 最后更新时间。

(5) 用户行为日志模型。用户行为日志模型来源于用户的显式行为和隐式行为。显式行为有: 我的收藏、用户注册、自助反馈。隐式行为有: 原文传递、代查代借、浏览、检索。

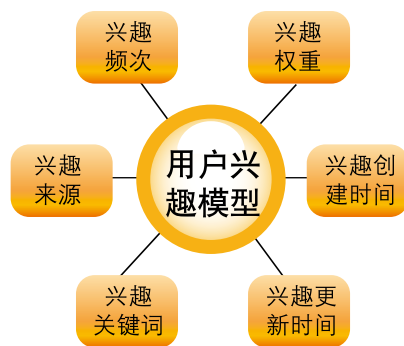


图2 用户兴趣模型

4 NSTL个性化推荐系统功能

在原有NSTL网络服务系统应用功能的基础上,借鉴国内外优秀文献数据库系统、电子商务网站的个性化推荐功能的设计,提出了NSTL个性化推荐系统的功能升级

设计。在原NSTL网络服务系统中的检索结果页面、命中文献页面、自助中心以及邮件系统中增加个性化推荐功能,实现NSTL的个性化信息服务。

4.1 基于群体用户的推荐

(1) 检索结果重排序。在系统检索结果默认排序基础上,为用户提供基于群体行为进行检索结果重排序的选择方案,包括两种情况:按点击量排序、按订购量排序。为了实现该功能提供了两种方案,一种是将群体行为信息加入到索引中,将重排序嵌入检索过程;另一种是对检索结果基于群体行为进行重新计算,不干预检索过程,命中量较大时代价较高,结果响应较慢。

(2) 共同行为推荐。该功能分为两种,一种为浏览此文献的用户还浏览过,另一种为订购此文献的用户还订购过。第一种功能基于用户浏览日志和NSTL文献资源模型。首先查找浏览过此文献的用户,再查找这些用户短时间内(前后10分钟)还浏览过的文献,通过文献标识号进行Group By计算,结果按浏览用户数倒排序;第二种功能基于用户订购日志和NSTL文献资源模型。首先查找订购过此文献的用户,再查找这些用户还订购过的文献,按订购用户数倒排序。

4.2 基于个体用户的推荐

该功能出现在用户登录后的自助中心,在“我的定制推送”的基础上,根据用户兴趣为用户推荐最新上线发布的相关文献(按期刊、会议录、文集汇编等类型分面),在用户自助中心提供推荐列表,也

提供邮件推荐。该功能基于用户兴趣模型和NSTL文献资源模型,按照用户权重较高、更新时间较新的Top N用户兴趣,与NSTL资源进行相似度计算,按相似度较高、文献更新时间较新的排序方式进行倒排序。

4.3 基于文献内容相关度的推荐

基于文献内容相关度的推荐包括以下两种场景:

(1) 相似文献推荐。该功能推荐与该文献内容相似度较高的文献,基于NSTL文献资源模型。首先根据单篇文献的关键词、标引词和题名等获取候选文献列表,进行相似度计算,按相似度较高、文献更新时间较新的排序方式进行倒排序。

(2) 相似作者和研究机构推荐。该功能推荐与该文献作者研究领域相同、合著共现的其他作者和研究机构,基于NSTL文献资源模型。计算与单篇文献作者研究兴趣(作者关键词、标引词等)密切相关、合著共现的其他作者;研究机构依据文献作者所属机构进行推荐。

5 系统核心模块设计与实现方法

5.1 同行为分析模块

此模型对应功能:浏览此文献的用户还浏览过;订购此文献的用户还订购过。

(1) 模块逻辑结构。该模块的逻辑结构如图3所示。

基本步骤:先查找浏览文献的共现ID,并通过此ID,查询所有文

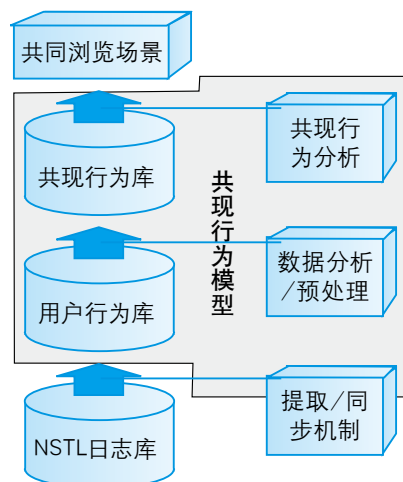


图3 同行为分析模块逻辑结构

献,可以利用数据库Group By计算,按浏览共现次数倒排。

(2) 数据分析/预处理。数据分析与处理方法为首先分析用户会话、浏览文献时间和文献所属分类,然后分析出共现行为,并把具有共现行为的浏览日志赋予统一的ID,即共现ID。

5.2 文献相似模块

(1) 模块逻辑结构。首先要获取当前浏览文献的主题、分类号。然后基于STKOS进行主题词映射,进而进行概念检索,将检索结果进行筛选(例如前20篇文献)并放入相似文本集合,将集合中的文献与浏览的文献就行相似度计算,从而获取最相似的文献资源。流程如图4所示。

(2) 基于STKOS相似度算法。该算法应用于用户兴趣概念与文献资源匹配阶段以及判断文献相似度的阶段。在系统中,我们要为每个用户的兴趣进行建模,其中兴趣概念是用户模型中的最重要输出,本模型针对兴趣概念进行建模。而这些信息最终都是映射

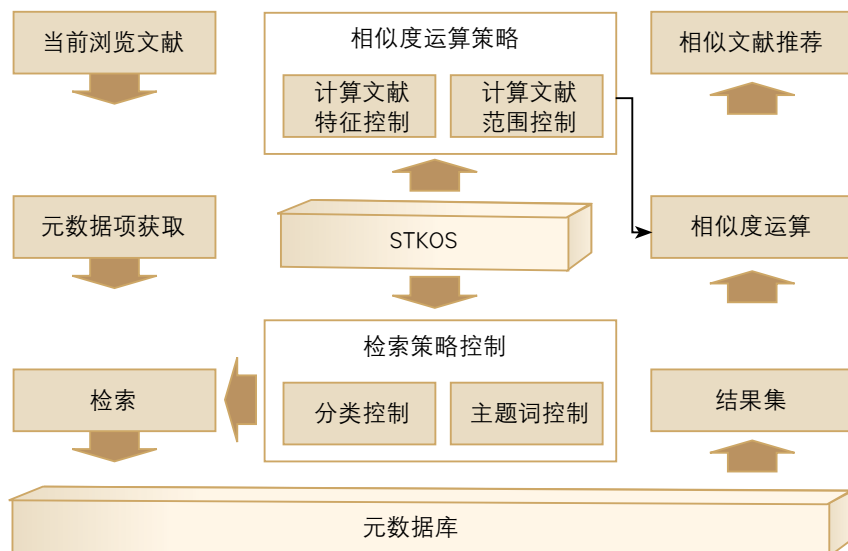


图4 文献相似模块逻辑结构

到STKOS的概念上,这样每个用户兴趣概念就是一个向量,该向量的每个分量就是主题词表中的词,

这样我们就可以通过计算两个向量的夹角的余弦来判断两个文献的相似度^[5]。

6 结语

作者对当前个性化推荐服务的应用情况进行了调研分析,并根据NSTL网络服务系统的个性化信息服务现状,为其进行个性化升级提出了总体结构、模型设计、功能设计方案。目前作者正在参照设计方案,着手进行NSTL个性化推荐系统的开发。

在未来的工作中,作者将进一步跟踪个性化推荐服务的最新动态,同时利用现有元数据集将模型构建起来,进行性能评测,最终应用到个性化推荐系统的开发之中,进而提升NSTL网络服务系统的个性化服务水平。

参考文献

- [1] 黄晓斌,张海娟.国外数字图书馆推荐系统评述[J].情报理论与实践,2010(8):125-128.
- [2] 熊馨.电子商务个性化推荐系统研究——个性化产品推荐策略研究及算法设计[D].中国科学技术大学,2005.
- [3] 秦凤蕊.个性化推荐技术研究及在数字图书馆中的应用[D].长春理工大学,2010.
- [4] 宋丽哲,唐赤兵,王胜海.基于本体的数字图书馆个性化用户模型表示[J].中文信息学报,2008(1):99-103.
- [5] 张玉峰,蔡皎洁.基于Web挖掘技术的用户兴趣本体学习研究[J].情报学报,2011,30(4):380-386.
- [6] Amazon [EB/OL]. [2012-05-12]. <http://www.amazon.com/index.html>.
- [7] 360BUY [EB/OL]. [2012-05-10]. <http://www.360buy.com>.
- [8] Dangdang [EB/OL]. [2012-05-15]. <http://www.dangdang.com>.
- [9] 万方数据知识服务平台[EB/OL]. [2012-05-20]. <http://g.wanfangdata.com.cn/>.
- [10] PubMed [EB/OL]. [2012-05-16]. <http://www.ncbi.nlm.nih.gov/pubmed/>.
- [11] CNKI [EB/OL]. [2012-05-17]. <http://cnki.caas.net.cn/kns50/>.

作者简介

胡小丰 (1988-), 硕士研究生, 研究方向: 信息管理与信息系统。E-mail: zac03725@hotmail.com

赵瑞雪 (1968-), 博士, 研究员, 博士生导师, 研究方向: 信息管理与信息系统、信息资源管理、知识组织及数字图书馆。E-mail: zhaorx@mail.caas.net.cn

Research on NSTL-Oriented Personalized Recommendation System

Hu Xiaofeng, Zhao Ruixue, Xian Guojian, Kou Yuantao / Agricultural Information Institute of CAAS, Beijing, 100081

Abstract: With the rapid growth of literature resources in digital libraries, the information service institutions ought to track the quality of personalized information services and to promote the efficiency of that, therefore the technology of personalized recommendation which is widely accepted by the users is developing increasingly. This paper introduces the present applications of the personalized recommendation in the information service patterns and E-commerce patterns around the world, and redesigns the NSTL internet system involving architecture, functions, key modules based on the current level of personalized information service in NSTL internet system.

Keywords: Information service, Personalization service, Recommendation system, NSTL

(收稿日期: 2012-08-01)