模糊语义个性化信息推荐

牟向伟 著

清华大学出版社

模糊语义个性化信息推荐

牟向伟 著

清华大学出版社
北京

内 容 简 介

个性化信息资源的建设组织过程是一项系统工程，本书在分析了个性化信息服务的内容、建设模式、目标和原则的基础上，提出了个性化信息服务体系和构建方法；阐述了在信息集成环境下信息资源的权限管理和访问控制方法，并建立统一的模糊语义模型来描述集成后的信息资源和用户特征。本书使用多种基于数据挖掘的方法对用户的兴趣进行深层次的发现，并对经典的协同过滤算法进行改进，使算法在保证准确度的同时提高了运行效率。最后，本书提出了一种模糊语义个性化推荐系统模型，并使用 FALC 模糊描述逻辑语言实现了该模型。

本书可以作为个性化信息系统与数据产品的设计与开发等相关技术人员的参考书，传统的互联网开发者、决策者和计算机相关研究人员也可以从本书中得到启迪。

图书在版编目(CIP)数据

模糊语义个性化信息推荐/牟向伟著. —北京：清华大学出版社，2017
ISBN 978-7-302-44997-3

Ⅰ. ①模… Ⅱ. ①牟… Ⅲ. ①模糊语言－研究 Ⅳ. ①H087

中国版本图书馆 CIP 数据核字(2016)第 216148 号

责任编辑：袁勤勇　李　晔
封面设计：傅瑞学
责任校对：梁　毅
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn,010-62795954

印 装 者：保定市中画美凯印刷有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：6　　**字　　数**：141 千字
版　　次：2017 年 2 月第 1 版　　**印　　次**：2017 年 2 月第1次印刷
印　　数：1～1000
定　　价：19.00 元

产品编号：070213-01

前言

随着网络时代的发展，网络上信息资源“量”的丰富和“质”的稀缺影响了信息化建设的效果，迫切需要一种能够发现用户内在需求，并主动提供信息服务的功能。近年来，在国内外兴起的个性化推荐成为解决这些问题的重要途径之一。个性化信息资源的建设组织过程是一项系统工程，为了满足社交网络、电子商务和政务等信息化应用的个性化服务发展的需要，本书使用多种基于数据挖掘的方法对用户的兴趣进行深层次的发现，并对经典的协同过滤算法进行改进，提出一种基于稳定度的协同过滤改进算法，通过选取用户维和项目维中对目标评价最有价值的、最稳定的信息，使得算法能够在“最近邻”或“最相似项目集”规模较小的条件下得到较高的准确度，使算法在保证准确度的同时提高了运行效率。最后，提出了一种模糊语义个性化推荐系统模型，其中包括了基于直觉模糊集的用户模型，它能够更加全面地描述用户兴趣的模糊信息，提高对用户兴趣的表达能力。例如，用户的“兴趣程度”、用户的“评价”和对兴趣的“犹豫程度”等。并使用模糊描述逻辑语言FALC实现了该模型，该模型可以在语义环境下为用户提供更准确的和扩展性更高的个性化推荐服务，有能力描述推荐系统中的模糊概念，如偏好程度、热门资源等。根据语义层面的描述以及概念层次上的继承信息，使得推荐结果被适当地扩展，在一定程度上解决了传统算法中的推荐结果多样性问题、相同特征词资源的相关性问题、冷启动问题和用户项目矩阵稀疏问题。

最后以某政府部门航务海事管理部门的个性化信息服务功能与需求为例，本书分析了电子政务类网站的个性化信息服务体系、建设内容、建设模式、目标和原则，提出了以电子政务门户网站为核心的个性化信息服务体系和构建方法，阐述了在信息集成环境下该类系统中信息资源的权限管理和访问控制方法，并建立统一的模型来描述集成后的信息资源和用户特征。在语义环境下为用户提供更准确的和扩展性更高的个性化推荐服务，使得电子政务信息服务类网站中的信息资源能被系统高效地吸收和利用，满足用户的个性化信息需求，发挥资源的最大效益，实现资源的合理配置和共享。本书的相关研究感谢“中国博士后科学基金资助项目(2014M551063)”“辽宁省教育厅科技研究项目资助(L2014203)”“辽宁省社会科学规划基金项目(L14BGL012)”“中央高校基本科研业务费专项资金资助(3132016046)”的资助支持。

作　者

2016年5月

目录

第1章 绪 论

1.1 研究背景与意义

随着网络时代的发展，爆炸性增长的网络信息资源使得用户和信息提供者都面临着“信息过载与信息饥饿共存”的尴尬局面[1]。一方面，用户在大量的信息面前无所适从，无法找到自己所需要的信息甚至有时无法意识到自己的真正需求；另一方面，信息提供者只能被动的提供信息，对所有用户提供相同的信息，并等待用户自己找到自己需要的信息。信息资源“量”上的丰富和信息资源“质”上的稀缺造成了各行业进行信息化建设时的资源浪费，最终严重影响了信息化建设的经济效益和社会效益。

近年来，个性化推荐已经成为解决这些难题的重要途径之一。国内学术界自2000年以来对个性化服务技术的研究逐渐成为热点。推荐系统能够为用户提供相关信息的推荐，帮助用户或者代替用户找到其最感兴趣的信息，从而使用户能够顺利高效地完成网上浏览和购买等过程，它使用户从被动的信息浏览者转变成为主动参与者。同时，信息提供者通过建立一个效果较好的推荐系统可以与用户建立更好的关系，为用户提供更好的服务，增加用户的忠诚度以及信赖程度，提高访问量。个性化推荐方法的总体思想是在用户的行为、偏好、历史记录中的信息以及其他相似用户的相关信息中挖掘用户的兴趣信息，并为用户提供个性化推荐服务，以达到“将最有用的信息推送给最需要它的用户”的目标。

个性化推荐系统已经在很多领域取得了巨大的成功，如个性化推荐技术已经应用在很多实际的电子商务系统中，国外的Amazon、eBay，国内的当当、淘宝和新浪等网站，都在不同程度上使用了各种形式的推荐系统，据VentureBeat统计，Amazon的35%的商品销售额由推荐系统为其提供。在这些电子商务系统中需要推荐的信息和用户的信息都是海量的，在日趋激烈的竞争环境下，电子商务推荐系统能有效提高系统对用户的“黏度”，提高用户的忠诚度，增加电子商务系统的销售额，发现用户潜在的消费倾向与偏好，为众多的用户提供个性化服务，它既能作为一种工具帮助用户过滤掉无用的冗余信息，又是一种网站营销的手段，帮助网站提高用户的忠诚度并推广各种相关产品或服务，在为用户提供便利的同时又为信息提供者带来巨大的经济效益。

在教育领域中，如美国的Illinois大学研究的CIRCSIIVI-Tutor项目能够支持远程网络教育的个性化服务，并通过建立一个基于语言对话框的智能系统，帮助学习者解决一定的问题[2]。国内上海交通大学的申瑞民教授研究了“基于数据挖掘的个性化学习导航系统”，该系统根据学习者的个人特征实现个性化服务[3]。电子教育领域中个性化服务以学习者为中心，根据学习者的学习能力、知识水平和兴趣爱好等特征为其推荐相应的学习内

容[4]，使学习者能够更加充分地利用网络教育和学习资源。

在电子政务领域，随着《国家信息化发展战略(2006—2020)》和《国家电子政务总体框架》出台，我国电子政务实施已经从基础设施建设阶段跨入面向社会化服务实施的新阶段。在此阶段，电子政务实施呈现出信息超载的加剧、用户信息需求的复杂性和差异性增加等多方面的特点。这些都要求为不同的用户提供个性化信息推荐服务。为了更好地实现实施电子政务的目的，更广泛地提高社会效益，个性化信息推荐服务将起到关键的作用。如新加坡政府的在线政府服务 eCitizen 网站，公民可通 My. Ecitizen 获得个性化页面、个性化提醒、个性化事项提醒等功能[5]。国内的青岛政务网为注册用户提供多种个性化定制服务，大连政府网还将用户划分为政府、市民和企业等，根据其类别的需求不同提供有针对性的个性化信息服务。

随着个性化推荐系统的发展，现有的方法的瓶颈越来越明显，如特征提取、冷启动、过拟合和稀疏问题等。尤其是 2000 年以来，在发明万维网 10 年之后，Tim Berners-Lee 提出了“语义网”的理念。使得万维网能够构建被计算机自动识别的语义信息标识，并使得计算机程序能够对资源(不仅限于 HTML 网页，也包括不能通过网络直接访问的对象)进行分析和推理。语义网的提出给个性化推荐系统带来了新的挑战，如何利用网络资源和用户历史记录中的“语义”信息来提供更好的个性化推荐服务已经成为个性化推荐领域的最新思路和重点研究方向。

1.2 个性化推荐系统概述

直到 20 世纪 90 年代末，个性化推荐系统才作为一个独立的研究领域被提出来，随着 Web 技术的发展和成熟，推荐系统在网络上的应用也蓬勃发展，国内外学术界对推荐系统的相关研究很多，最有影响力的个性化服务系统是由 Stanford 大学提出的协同推荐系统。2000 年以来，在个性化服务相关技术逐渐成为国内的研究热点，很多国内学者和研究机构已经开发出了一些原型系统，某些网站也推出了简单的个性化推荐服务的功能。openBookinark 是清华大学推出的一个混合推荐系统，它的混合推荐是通过集中管理用户群的 Bookmark 来实现的；DOLTRIAgent 系统由南京大学潘金贵等人设计，该系统实现了个性化信息检索智能体；由上海理工大学的陈世平等研究和开发的个性化智能检索系统 MySpy 利用辅助词典、同义词词典和蕴涵词词典来实现基于智能代理的信息过滤和个性化服务，对查询词进行概念扩展，并给用户返回与查询需求相似的文档；万方数据开发的 iLib 系统拥有针对相似资源推荐的功能，该功能可以推荐与用户当前访问的文献资源具有高相似性的其他资源；中国期刊全文数据库(CNKI)，不仅可以提供相似资源的推荐，还可以根据科技学术文献的被引文献、同作者文献等引用信息进行推荐。推荐系统涉及的领域广泛，包括信息管理、信息集成、数据挖掘、知识管理、信息检索、预测方法等。根据推荐系统的使用方法不同，可以将推荐系统分为以下几类：

(1) 基于规则的推荐方法，使用 if-then 的形式定义知识规则。如，在 WebSphere Personalization Solution 中，可以定义“规则”使不同的内容显示给特定的用户。Broadvision 使用专家规则(Expert Rules)来实现个性化商务服务[6]。

(2) 协同过滤推荐方法[7]通过最近邻(具有相似兴趣的用户)或相似项目(相似度最高的项目)的意见或评分来预测某用户对某项目的意见或评分。协同过滤系统已经得到广泛应用,并取得了巨大成功。其推荐过程可以分为两步:一是利用用户的历史信息或待推荐项目的评分信息计算用户或项目之间的相似程度;二是利用与目标用户或目标项目有相似程度较高的用户或项目的评分(评价)信息来估计目标用户目标项目的偏好程度,系统根据这一偏好程度来对目标用户进行推荐。协同过滤推荐系统的核心是根据用户和项目间的评分或评价信息来产生"用户-项目"矩阵,当新用户出现在系统中的时候,还没有对任何项目进行评分,系统无法挖掘用户的偏好,因此系统无法使用协同过滤推荐一类的方法给新用户提供准确的推荐服务;同样,新的项目在推荐系统中出现的时候,还没有任何用户对该新项目进行评分,而协同过滤系统必须完全依靠评分或评价信息进行推荐,所以直到新加入的项目被一部分用户评价,系统根据这些信息才有可能产生针对此项目的推荐。很多研究称此问题为冷启动问题,现在有很多方法在一定程度上解决了这个问题,例如利用启发式算法度量用户之间的相似性或者在"用户-项目"矩阵中提前加入伪打分信息。另一个问题是,在任一实际应用的推荐系统中,通常已经评分项目的数量很少,从这些少数的项目中很难产生准确的推荐,这个问题称为稀疏问题,可以用辅助信息获取和信息扩散方法解决该问题,另外一个解决方法就是在计算相似性的时候可以使用用户配置文件(User Profiles)中的历史记录,而不仅仅考虑对商品的评分。面对日益增多的用户,虽然协同过滤推荐算法能利用最新的用户评价为用户产生及时的、相对准确的推荐,但是随着系统中用户和项目数量的急剧增加,用于发现用户的"最近邻"(相似用户)和待推荐项目的"最相似项目"算法的计算量也大大增加,对于拥有上百万用户的系统,算法通常会遇到扩展性瓶颈问题,严重的会直接影响到推荐服务的实时性和准确性,如何适应系统规模不断扩大的问题成为制约系统实现的重要因素,该问题称为扩展性问题。目前解决扩展性问题的主要方法是通过机器学习提前进行训练,这种方式被称为基于模型的协同过滤算法,虽然基于模型的算法在一定程度上解决了扩展性问题,但是由于模型的学习过程以及模型的更新过程,在算法的及时性和最新信息的利用率上该算法要差些,同时该算法的准确性也不及协同过滤算法。所以该算法仅适用于用户偏好较稳定的情况。基于协同过滤的推荐系统包括 Tapestry[8]、GroupLens Project[9]、WebWatcher[10]等。

(3) 基于内容的推荐方法通过分析项目的内容和描述来确定用户的兴趣[11]。基于内容的推荐方法来源于信息过滤,其中有很多方法都是相似的。所谓"内容",可以包括网上各种形式的数据,可以是网页、多媒体、数据项、电子商品、用户行为记录和服务器日志等。基于内容的推荐方法摆脱了用户对项目的评分信息的依赖,它利用用户浏览或查询的项目的具体内容信息来建立用户和项目之间关系,然后产生相应的推荐。基于内容的推荐系统需要对用户和产品分别建立描述文件,借助数据挖掘或机器学习方法对用户已经购买、浏览或查询过的内容进行分析,据此建立或更新用户的兴趣描述文件。通过用户与项目兴趣描述文件之间相关程度的比较,最终将与用户兴趣描述文件最相似的项目推荐给用户。基于内容的推荐方法所用到的信息获取和信息过滤是算法的核心技术,在基于内容的推荐系统中,用户或项目描述文件中的特征需要被计算机识别并提取,但是图形、视频、声音文件等多媒体数据在特征提取的时候存在一定的困难。同时,由于语义信

息的缺乏，可能造成同一个特征词同时存在于两个不同的项目描述文件中，算法根据这个特征词无法区分两个项目的不同，这类问题需要加入语义信息等工具来解决。另外，基于内容的推荐算法中用户得到的推荐项目是那些只与用户的描述文件相关度较高的项目，推荐的结果是与用户之前偏好项目相似的项目集，推荐的内容会被局限在一个特定的范围内，无法保证推荐的多样性。现有的基于内容的推荐系统包括 ELFI[12]、CiteSeer[13]、Personal WebWatcher[14]等。

(4) 基于网络结构的推荐算法把用户和项目看成抽象的节点，将用户与项目之间的关系抽象成链接节点的边，算法利用的信息都隐含在用户和项目组成的“用户-项目”网络之中，该算法的核心是要建立用户商品二部图关联网络。基于网络结构的推荐算法同样受到冷启动问题的制约，新用户或新项目出现在系统时没有任何选择或被选信息，用户与项目的关系网络无法建立，推荐结果也就无从产生，如果考虑新用户或项目的描述文件中的内容或者利用启发式算法可以在一定程度上解决此问题。当用户兴趣更新时，需要重新遍历“用户-项目”才能得到新的推荐结果，导致推荐算法的效率降低。另外一个值得关注的问题是如何在基于网络的推荐中加入历史时间的影响。

以上几种方法所使用的关键技术如表 1.1 所示。

表 1.1 推荐算法关键技术

推荐方法		关键技术方法	面临的问题
基于规则的推荐方法		事件动作规则，逻辑语言，专家系统，规则引擎	缺少自动化处理过程，需要管理员手动添加规则
基于内容的推荐方法		tf * idf，向量空间模型，文本分类，机器学习等	无法保证推荐的多样性，无法自动识别多媒体数据，特征词无法区分不同的资源
协同过滤推荐方法	基于用户	最近邻算法，预测算法等	冷启动问题，可扩展性问题，稀疏问题
	基于项目	相关度计算方法，预测方法等	
	基于模型	聚类，机器学习等	
基于网络结构的推荐算法		网络遍历算法，网络结构等	更新效率低下，无法体现历史事件的影响

上述的推荐算法都面临着一些共性的问题。例如，当新用户和新项目加入、用户最新动态更新的时候，用户和项目的信息需要进行动态更新，如果每次更新后都需要重新计算，将耗费巨大的时间和计算资源，因此需要定期进行脱机计算，使系统的在线服务尽量不受影响。另外一种解决方法是设计某种近似的动态算法，每次针对局部更新结果改变原来的算法结果，避免不必要的完全重新计算。上述各种推荐方法都有优缺点，所以有很多研究采用混合推荐策略取长补短，其做法是首先选择两个或多个推荐方法产生一个推荐预测结果集，之后通过某种方法对结果进行组合后反馈给用户。另外，针对推荐系统已经有很多评价指标被提出来，包括准确率、召回率、平均打分值、商品平均度、差异性等。在不同领域应用的推荐系统使用不同的评价指标得到的结果是不同的，利用不同数据集产生的推荐结果也是不一样的，因此推荐系统研究的另一个重要问题是如何选择合适的

评价指标对推荐效果进行评判。

目前，在语义网环境下建立能够处理语义知识的个性化推荐系统已经成为该领域的热点问题，基于万维网创始人 Tim Berners-Lee 提出的语义网的概念，个性化推荐系统除了要能够根据用户和项目的评分和内容信息进行推荐，同时还要能够根据其中隐含的语义信息进行推荐。例如，通过给某个用户的兴趣加以语义标注，就能够区分“JAVA”这个兴趣是属于“计算机语言”还是“地理信息”。借助语义信息的推理，推荐系统还能够对用户的兴趣和资源的描述进行适当的扩展，发现用户兴趣中的隐含概念，丰富用户的兴趣信息，进而改善推荐结果。相关研究包括 Szomszor M 研究了如何利用语义网环境下的语义信息进行电影推荐[15]；Schickel-Zuber V 等提出了一种可应用在推荐系统中的本体学习方法[16]；清华大学的梁邦勇等人提出了基于语义的网页推荐模型[17]；董兵、吴秀玲对个性化知识推荐系统的语义扩展问题进行了研究[18]。

综上所述，虽然基于语义的个性化推荐系统已经成为该领域的主要研究方向之一，但是相关研究比较少，尚处于起步阶段，而且在推荐系统中存在大量模糊概念没有受到足够的关注，如用户对于资源的偏好程度，资源与某个概念的关联程度，以及热门资源、受欢迎资源等都是模糊概念。建立能够处理模糊语义信息的个性化推荐系统不仅可以提高推荐系统的效率，还能改善用户体验，增强推荐系统对用户兴趣和资源描述的表达能力。

1.3 管理信息系统对个性化推荐服务的需求

信息化建设是管理决策部门高效行使职能、履行职责的有效手段，管理信息系统可以打破组织结构限制、时间限制和地域限制，可以为用户提供有效便捷的沟通与服务的平台，同时也是各级组织机构内外部交流沟通平台。管理信息系统涵盖了办公自动化、信息维护、业务流程管理以及内外部沟通交流等多种功能，其特点是应用范围广，内容丰富，实时互动。信息管理系统的实施可以起到以下两个重要作用：

(1) 满足建设高效服务的需要。有利于大幅度改进管理决策部门的工作效率和服务水平，规范管理行为、整合跨部门职能、推进机构改革、扩大服务范围。

(2) 更好地调配、调节、引导信息资源，为业务和客户服务。管理信息系统中拥有组织机构中最全面、最权威的信息，信息的规范使用可以提升管理水平，加强科学化管理，提高管理工作效率。尤其在电子政务方面，发展初级阶段，对于如何为社会提供优质的服务还存在不少问题。信息化领导小组的咨询专家、著名的经济学家吴敬琏指出，目前我国电子政务存在的问题是重新建轻整合、重硬件轻软件、重管理轻服务、重电子轻政务[20]，其主要问题还包括：

① 服务方式有限。目前的信息服务还停留在发布一些简单的静态信息，如法规、指南、政府机构介绍等，没有考虑到民众对信息的个性化需求，信息服务缺乏主动性和互动性。服务方式仍旧是“被动服务”、“人找服务”的服务方式。

② 服务功能简单。民众虽然可以从电子政务网站上了解一些政务信息，但还是缺乏在网上办理事务的必要渠道。网站内容还是按照现实政府机构排列，没有按照民众的需求排列，要得到跨部门事务处理的信息，还需要逐一进入各个部门查询，没有考虑到民众

操作的便捷性。

③ 条块分割，信息孤岛现象严重。电子政务的建设处于各自开发、各自为政的状态。缺乏统一的标准，资源无法得到共享，这种情况不仅造成重复建设和资源浪费，也给民众的信息获取带来一定的困难。

以上这些问题不仅使民众感受不到政府电子化服务的优越性，而且也会造成民众参与的积极性降低，最终影响电子政务的建设。随着网络时代的发展，信息技术的进步，电子政务上所包含的信息量将越来越多，而民众的需求也是个性化的，单一的服务模式很难满足特定单一民众或某个群体的具体需求。同时，面对专业性强、信息量大、知识结构复杂的信息，民众无法判定自己需要哪些信息，如政策法规、航务海事、交通物流等，都需要一定的专业性，缺乏相关知识的用户在获取知识时往往无所适从。因此，民众迫切需要一种能够自动发现他们特定的个性化需求，并根据这些需求来组织和调整信息的服务，即个性化推荐服务。本书在分析了电子政务个性化推荐服务的体系、建设模式、组织管理方法的基础上，提出了模糊语义个性化推荐系统模型，它可以在语义环境下提供个性化推荐服务，使得电子政务中的信息资源能被用户高效的吸收和利用，还可以满足民众的个性化信息需求，发挥资源的最大效益，实现资源的合理配置和共享。

第2章

模糊语义个性化推荐系统理论基础

模糊语义个性化推荐系统的理论基础包括了信息集成理论、数据挖掘技术和模糊描述逻辑。其中信息集成理论帮助推荐系统收集不同来源的异构信息，整理用户的行为记录。数据挖掘技术负责在集成的信息中发现用户的兴趣，挖掘待推荐信息资源中的模式和知识，最终产生推荐。模糊描述逻辑为推荐系统提供了处理模糊语义的能力。

2.1 信息集成与管理理论

信息技术的发展为社会带来了前所未有的变革，它是继工业革命后的又一次技术飞跃[21]。随着计算机互联网和通信技术的快速发展，以信息技术的大规模发展、渗透、扩张和利用为基本内容的社会信息化活动已经成为推动一个国家和社会发展的最活跃的因素之一[22]。信息技术的飞速发展已经影响了社会各行业的环境，并将持续发生更加深刻的变化，只有适应这种变化，才有生存和发展的空间。在过去的30年中，不同行业不同领域都进行了不同程度的信息化建设，在这些早期的系统中，信息往往只支持业务过程系统的独立性、离散性，难以体现各环节之间的关系，形成信息孤岛或信息断层，造成企业生产经营、决策过程的堵塞和不联系性[23]。信息的管理者和使用者都面临海量的、分布的、异构的信息。人们获取相关信息的能力在信息化进步的今天反而更显艰难和无奈，因此人们开始关注如何将不同环境中的异构的信息资源集成化，不仅要提供信息资源的集成环境，而且要能够提供更加友好的用户使用环境。

2.1.1 信息集成理论

在众多复杂的系统中，将浩瀚的信息进行集成需要通过一定理论来指导具体的实践[24,25]，国内外指导信息集成的基础理论包括系统论、信息集成原则、信息集成模式、知识组织理论等。

(1) 系统论。系统论把对象以系统的形式加以观察，以系统的角度指导信息化建设实践过程，从关联性、整体性和优化性进行考察。使得由各具体资源整合而成的信息的集成体系以系统论为指导。系统论作为理论基础具有重要的现实意义[26]。

(2) 知识组织理论。知识组织理论旨在揭示知识的本质和知识间关系；知识组织通过元数据格式对信息进行描述，整合异构数据，以实现不同资源和系统间的资源共享，并发掘具有内在关联的信息链、知识链和知识内涵；优化知识库结构，以加强知识利用和创新能力；在知识发现技术的基础上，知识组织可以实现更多功能，如提取、转换、过滤、整合

等对异构数据的操作；在智能知识抽取和处理过程中，信息资源按特定的方式表示并以知识内容特性进行聚集等。不论从何种角度（技术、形式、组织对象、组织方式）来看，所有一切都表明数字资源整合应该建立在知识组织理论的基础之上[27]。

（3）信息资源整合过程的指导原则包括：保证资源集成的发展性和不间断性的连续性原则；保持资源对象学科的完整性和整体性原则；强调集成的目的是满足特定用户的需求的针对性原则[26]；运用技术手段和方法优化组织结构和功能的优化性原则；强调集成的结构性和多维性的层次性原则；针对集成对象、内容、方式的科学性的科学性原则[28]。

（4）信息的集成模式包括：关联模式按信息内容间的相邻性将有关信息集成在一起；组织模式使用结构特性将信息组织在框架内结构；综合模式将相关内容从信息中提取出来并重新组织为新的信息；分析模式对原始信息进行分析并利用一系列定量或定性分析模型得出结论性或咨询性信息[29]；基于数据仓库与数据挖掘的信息集成模式在数据仓库的基础上，利用知识发现技术、数据库转换技术和基于多平台异构数据整合方法与标准，为高层管理提供决策支持[21]。此外，针对图书馆资源的集成提出的多元集成模式，如CNKI的完全集成式；中国数字图书馆的元数据集中、对象数据分散的集成式；以网络虚拟方法连接各信息资源进行数字化信息资源建设、管理、服务为主要任务的集成式；以各单位信息资源建设为主的集成模式[30]。

还有些研究探讨了相关因素的集成问题。如：从宏观环境的角度出发，提出的基础设施、应用软件和信息标准的三位一体信息集成环境[31]。这个环境应该是交互的、开放的、柔性的、动态有界的，并具有良好的组合、公共、互操作、兼容、可扩展等特性。不同部门逐步地分别地对异构或异质的信息资源进行描述、组织、开发和管理；从微观环境角度出发，集成环境或集成标准化问题是由信息加工、分析工具和用户服务界面三者有机结合组成的[29]。信息集成是资源开发、信息资源组织、信息管理的重要目标，并且实现这一目标的关键是标准化；信息集成的重要环境因素还包括了人的主观因素、集成系统的结构等。信息应该被看作是一种战略资源，我们应该以重视需求、系统地、创新的可持续发展观念进行系统集成，在进行系统集成的同时还要进行相对应的改进管理机制，改善服务结构，并进行人员和相关业务的调整[32~34]。

2.1.2 信息集成技术方法

近年信息集成技术方法研究比较侧重于系统集成的分布式服务构架、智能化及自动化方法[35~37]。最新的研究热点内容包括面向Web服务的SOA（Service Oriented Architecture）信息集成框架模式、基于Ontology本体论的信息集成方法和基于Agent理论的信息集成方法和中间件技术等。

（1）面向服务的信息集成框架模式（Service Oriented Architecture，SOA）。SOA是一种利用组合Web Service进行分布式应用集成的架构，SOA服务架构的基础是各种业界的标准规范，如OASIS（Organization for the Advancement of Structured Information Standards）、W3C（World Wide Web Consortium）和OGC（Open Geospatial Consortium）的Web服务相关规范。其中OASIS的BPEL工作流给出了关于组合服务的规范指导，OGC是一种互操作规范，包括CSW、WFS、WCS、WMS、WPS等，对数据及其元数据的服

务协议制定了规范，W3C 的 SOAP(Simple Object Access Protocol)协议与 WSDL(Web Services Description Language)协议是 Web Service 的基础协议。此外，在上述基础标准规范的基础上，国际标准化组织提出了系列补充协议以适应 SOA 智能化发展的要求，补充协议包括 Web Authority Service、Web Service Policy、Web Service Addressing 与 Web Security Service。

(2) 本体论(Ontology)。信息表达上的语义异构是由于描述信息没有采用统一的语法描述格式造成的，系统中语义异构的主要表现如下：不同的信息源中同一术语表达不同的含义；多种术语在不同的信息源中表示同一概念；一些概念间的隐含联系由于各信息源的分布自治性而不能体现出来。信息集成要解决系统间信息在系统间交换和理解的问题，实现包括信息的统一表示与信息转换以及基于信息理解的智能化检索等。目前本体被认为是解决语义集成的有效的手段之一[38,39]。基于本体的信息集成研究始于人工智能及知识工程领域，主要解决知识重用和共享问题[40~42]。目前的应用研究有：Stanford 大学的 SKC(Scalable Knowledge Composition)，解决了信息系统(包括 Web)中的语义异构问题，并实现异构系统的互操作；Ariadne 项目着眼于开发能够抽取、查询和集成 Web 信息源的智能 Agent；Observer 项目使用不同的本体来表达不同的信息源，并建立本体间的映射集合。Picsel 系统定义了一个基于知识中间层来连接用户和相同领域内的若干信息源，处理用户的查询并将查询结果返回给用户[43]。

(3) Multi-Agent System(MAS)以 Agent 理论为基础，注重系统集成行为研究。其原理是：Agent 成员并不能限制其他 Agent 的目标和行为，Agent 相互之间的矛盾和冲突通过竞争和磋商等手段来解决，因此 Agent 个体不能够解决的大规模复杂问题可以通过 Agent 团体的交互式协调来求解[44]。通过 Agent 个体以及群体的活动规则的建立来提高系统的智能化水平和适应环境的能力。Jennings[45]等人开发了一个基于 Agent 的集成框架 ADEPT，将各个子系统视为一个个智能代理，系统集成是通过这些智能代理之间的交互来实现的。

(4) 中间件技术。中间件是一种独立的系统软件或服务程序，分布式应用软件借助这种软件在不同的技术之间共享资源。中间件位于客户机/服务器的操作系统之上，管理计算机资源和网络通信，它是连接两个独立应用程序或独立系统的软件[46]。相连接的系统，即使它们具有不同的接口，但通过中间件相互之间仍能交换信息。执行中间件的一个关键途径是信息传递。通过中间件，应用程序可以工作于多平台或操作系统环境。

最早具有中间件技术思想及功能的软件是 IBM 的 CICS，但由于 CICS 不是分布式环境的产物，因此人们一般把 Tuxedo(1984 年由贝尔实验室开发完成)作为第一个严格意义上的中间件产品。IBM 的中间件 MQSeries 也是 20 世纪 90 年代的产品，它的许多中间件产品也是在近几年才作为成熟的产品来使用的。中国的中间件软件产品起步较早，与国外技术差距不大。如：北京东方通科技发展有限责任公司 1993 年推出第一个产品 TongLINK/Q，与 IBM、Oracle 在我国市场形成三足鼎立的局面，根据赛迪顾问、计世资讯、易观国际等咨询机构的市场分析报告，东方通中间件的市场占有率在国内企业中名列首位。在国内的科研院校中，中科院软件所早在 1995 年就开始利用“对象技术中心”的技术基础研究中间件。与此同时，国内还有国防科技大学、北京航空航天大学等研究机构也

对中间件技术进行了同步研究。

2.2 数据挖掘技术与方法

2.2.1 数据挖掘技术

有关用户和资源的个性化信息的抽取是个性化推荐过程中的关键步骤，最常用、最有效的方法是使用数据挖掘技术在历史数据中发现个性化信息。面向个性化服务数据挖掘的对象包括用户的历史数据、浏览模式、查询结果、服务器日志等，目标是最终发现用户的特定需求并且与可进行推荐的资源进行匹配。

数据挖掘是从大量的、不完全的、有噪声的、模糊的和随机的数据中挖掘潜在的信息和知识的过程，它是一种基于事实和数据的寻找对决策支持有用的模式的过程。与数据检索、查询不同，它需要对数据进行统计分析，加以综合和推理，发现事物间的相互关联，并利用已有的数据对未来活动进行预测。数据挖掘方法可以分为两类：一类是建立在统计模型的基础上，采用的技术有决策树、分类、聚类、关联规则等；另一类是建立在一种以机器学习为主的人工智能模型的基础上，如遗传算法、蚁群算法和神经网络等。

2.2.2 Web 挖掘技术

Web 挖掘是数据挖掘技术在 Web 环境中的使用。Web 数据挖掘的主要目的是自动从 Web 文档或服务使用记录中获取有用信息[47]。Web 使用挖掘、Web 结构挖掘和 Web 内容挖掘是 Web 挖掘的三个主要类别[48]。而语义 Web 挖掘则是当前研究的前沿领域。

(1) Web 结构挖掘主要分析页面之间的关联信息以及页面质量和结构等方面的特征，挖掘文档之间的引用、包含和从属关系，从 Web 超链接结构中发现某些模式，可以帮助查询检索提供更准确、覆盖面更广的结果[49]。比较流行的算法如 HITS[50]、Page Rank[51]等，这些算法主要用作计算每个页面的质量和相关性的手段来给出模型化 Web 拓扑结构。另外，还有许多研究集中在 XML(eXtensible Markup Language)文档的结构模式上[52,53]等。

(2) Web 内容挖掘，将网站内容分类，将类似网页组合在一起以提高信息检索的性能[54]。Web 内容挖掘主要分为 Web 文本挖掘和 Web 多媒体挖掘。Web 文本挖掘是从大量 Web 文本文档的集合中发现某种隐含的模式过程。为 Web 文本内容建立特征模型是 Web 文本挖掘中的关键技术，Web 内容挖掘中的聚类、分类、规则或模式识别等任务都与内容特征模型相关。常用的文本内容特征模型包括布尔模型(Boolean Model)、聚类模型(Cluster Model)、基于知识模型(Knowledge-Based Model)、概率模型(Probabilistic Model)和向量空间模型(Vector Space Model)等[55]，必要时还需要在特征进行提取与缩减[56,57]之后再利用关联规则、分类和聚类等数据挖掘方法提取规则模式。最后评价挖掘结果并分析改进接下来的挖掘工作。Web 多媒体挖掘是从 Web 上大量的多媒体数据(音频数据、视频数据和图像数据)中发现隐含的模式。相关应用如多媒体信息检索、多媒体信息建模、分类预测分析、多媒体关联分析等[58]。