

非中心化的机构知识库系统构建^①

张明宝, 李 雨, 扈 钰

(南京航空航天大学 经济与管理学院信息管理与电子商务研究所, 南京 211106)

摘 要: 中心化的机构知识库构建策略难以满足用户多样化和个性化的知识应用需求. 提出基于非中心化策略构建机构知识库. 设计了其系统结构和工作过程, 并对其关键技术问题作了概要阐述. 重点围绕 KM@web2.0 系统介绍了解决这些关键问题的实现方案.

关键词: 机构知识库; 非中心化; 情报; 知识管理; 协作

Decentralization Strategy Applied in Building Institutional Repository System

ZHANG Ming-Bao, LI Yu, HU Yu

(Department of Information Management and Electron Business, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: The center strategy of building institutional repository is difficult to meet the diverse and personalized needs of knowledge applications. It is proposed to build institutional repository based on decentralization strategy. Its system structure and working process are designed, a summary of its key technologies are elaborated. Some solutions to such critical issues are introduced and practice has proved that the decentralization strategy is feasible to build institutional repository.

Key words: institutional repository; decentralization; information; knowledge management; cooperation

机构知识库(IR)是指大学或研究机构为方便其成员组织、整理、存储和利用各类相关的数字知识资源而提供的一系列知识管理服务^[1]. 机构知识库的内容包括数字化的会议论文、论文预印本、研究报告、学位论文、技术报告、学术单位的新闻通讯和快报及档案、项目申请报告、图片文件、声像文件、数据集、课件等. 目前, 国外有代表性的机构知识库及其项目^[2]有: DSpace 联盟工程、加拿大 CARL 的机构库试验项目、eScholarship 知识库. 我国的机构知识库项目有香港大学的 HKUTO、浙江大学的 Dspace 系统、厦门大学的机构存储系统等等.

机构知识库的理论研究主要聚焦于机构知识库的设计理念、服务模式、版权问题、激励机制等方面. 文献[3]讨论了机构知识库发展中存在的版权许可、同行评议和质量控制问题、长期保存问题、标准化问题、内容收集问题、基金和持续性问题以及职员工作量的

潜在增加问题. 文献[4]探析了机构知识库如何在网络数字环境下为大学图书馆提供新的资源及服务拓展空间. 文献[5]提出了高校图书馆构建知识库过程中应采取的五项措施. 文献[6]提出了一系列保证机制来实现机构知识库的运营成功. 文献[7]从设计理念、技术平台、内容管理和服务功能方面对机构知识库的两个软件系统进行了分析比较, 并对未来该类软件系统的发展趋势做出展望.

机构知识库是数字化图书馆以及知识库应用发展等领域的自然延伸. 现有机构知识库研究主要聚焦于其应用与管理, 但是对机构知识库的构建机理和技术手段缺乏研究. 信息技术的发展经验告诉我们, 技术和工具是推动应用发展的前提和基础. 机构知识库系统已经成为机构知识库研究领域的重点问题. 现有的机构知识库系统研究^[8,9]都有一个基本假设, 即机构知识库是中心化的. 中心化的机构知识库易于管理和构

① 基金项目: 国家社会科学基金(13BTQ051)

收稿时间: 2014-07-08; 收到修改稿时间: 2014-10-13

建,但是在知识应用的多样性和灵活性方面却面临诸多挑战. 本文针对中心化机构知识库存在的问题提出一种基于非中心化策略的机构知识库构建方法,并且对该方法所涉及的关键问题和技术方案作了详细的讨论.

1 非中心化的机构知识库系统应用场景

传统的机构知识库系统无一例外都是基于中心化的策略构建,即机构知识库系统由机构的组织者负责构建,由它来搜集、组织、存储和发布各类知识,并且提供使用这些知识的工具和手段. 中心化策略最大的优点在于易于实现,但是所有知识都需要由知识库管理者来提供,很难实现海量知识供给和灵活面向用户需求这两大目标.

所谓非中心化策略是指在一个知识库管理者的协调下由广大机构知识库的用户来搜集和发布知识,知识库管理者负责组织、存储和管理这些知识,由用户和管理者共同提供知识应用的各种手段和方法. 非中心化的策略强调通过用户的广泛参与构建知识库. 在非中心化的机构知识库系统中,知识由广大用户提供,机构知识库的管理者变成一个协调者,由它来组织和协调用户贡献知识和描述知识的过程. 该策略最大的优势在于面向用户需求,适应环境变化,并且能够实现机构知识库的跨组织的集成与协作,它能为海量知识库的构建奠定坚实的基础.

非中心化的知识库系统构建最大的问题在于知识表示的无序,而要使知识表示有序则必须要借助于知识表示规范或者知识表示元数据. 传统上,知识表示标准是由某一权威组织发布和维护的,其构成策略也是“中心化”的. 中心化的知识表示规范的构建难度非常大,并且知识表示规范一经建立就很难改变,缺乏必要的灵活性^[10]. 为了解决这一问题,本文借鉴 web2.0 中社会化标签的发展思路,强调通过用户自我发展知识标签,然后利用大家共享协作构建的标签来描述和说明自我提交的知识单元. 我们将知识描述单元和知识单元分别存入机构知识库中的知识描述数据库和知识存储数据库. 在此基础上,用户就可以通过对规则化的知识描述数据库的检索和分类导航快速发现自己所需要的知识单元. 图 1 所示为我们提出的非中心化机构知识库系统结构图.

图 1 所构建的应用场景中包含两类角色,即知识

库管理者和知识库用户. 知识库管理者利用其所制定的机构知识库规范来协调用户在表示知识、共享知识和应用知识时所遇到的问题. 机构知识库管理者需要管理三类知识库,即知识存储数据库、知识描述数据库和知识标签数据库. 虽然这三个数据库中的内容都来自于用户,但是,我们所提供的机制对这些内容都进行了一定程度的规范化处理.

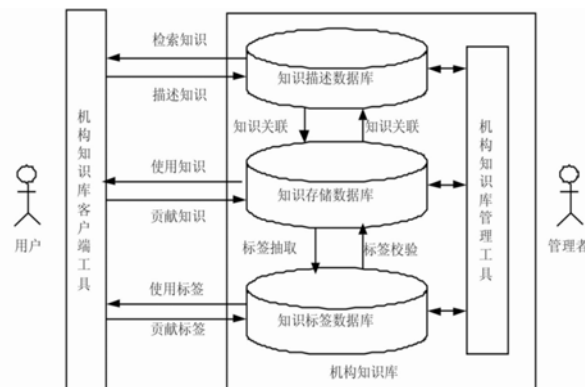


图 1 非中心化的机构知识库结构图

图 1 中的用户既是机构知识库的用户也是机构知识库中知识的贡献者. 为了使用户贡献的知识单元易于组织、管理和应用,用户必须对其所贡献之知识进行详细的描述和说明. 用户要贡献知识,首先必须要描述其知识,形成该知识单元对应的知识描述文档. 该知识描述文档应该遵循规范化的语法格式和明确的语义标准. 规范化的知识单元描述文档被组织起来存入知识描述数据库,成为未来用户检索和发现知识库中资源的关键依据. 知识单元描述文档是对相应知识单元的规范化描述,它是用来自于知识标签数据库中的知识标签来标注知识对象形成的. 区别于中心化策略,知识标签数据库中的知识标签也是由用户遵循一定规则通过协作共享构建的.

下面从用户的角度来完整描述非中心化的机构知识库系统中的完整知识表示、知识共享和知识应用过程.

用户首先形成了自己的知识单元,他需要将其递交到机构知识库共享. 在向机构知识库递交该知识对象之前,用户对该知识对象进行描述,形成知识描述文档. 在描述知识对象的过程中需要使用知识标签数据库中的知识描述标签. 用户上传其知识单元存入知识存储数据库,同时用户上传其对应的知识描述文档

存入知识描述数据库。知识描述文档与该知识单元之间建立了一一对应的知识关联。

当用户需要使用知识存储数据库中的知识对象时,它是通过对知识描述数据库中的知识描述文档的检索和分类目录导航发现相应的知识描述,然后通过知识关联取得具体的知识单元对象。由于知识描述文档使用的知识标签属于受控标签,所以,基于这些标签可以实现有效的信息检索和分类目录导航。

知识标签数据库中的标签也是由用户贡献的。它既可以来自于用户的有目的的设定,又可以来自于用户标注文档过程中所产生的随机性标签,还可以来自于使用数据挖掘技术从知识描述数据库和知识存储数据库之间的知识关联中分析所得。为了简化用户的使用过程,知识标签的采集主要借助于知识库客户端工具来完成。

图 1 所示的非中心化机构知识库系统的特点在于:1)由广大用户参与贡献知识,能够有效实现机构知识库的海量知识获取;2)由用户贡献的知识单元是被自我描述的,并且其描述文档是一定程度规则化的,能够实现机构知识库中知识的有效挖掘和利用;3)用于描述知识的标签也是用户贡献的,是一定程度受控的,能够为知识表示的规范化奠定坚实的基础。

2 非中心化的机构知识库系统构建的关键问题

社会化标签的发展一方面展示了通过社会大众参与协作实现信息共享的巨大发展空间,另一方面也揭示了社会标签无序化所带来的信息共享缺乏深度的严重问题。图 1 所示的机构知识库系统构建方法正是在借鉴社会化标签发展的经验教训的基础上提出:一方面,我们强调通过大众参与构建机构知识库的内容,另一方面我们强调利用受控标签来规范化表述用户贡献的知识。该方法区别于传统方法的最大特点在于其所使用的知识描述标签也来自于社会大众的参与共享。

上述策略面临一个突出矛盾,即大众参与强调简单,而受控知识表示又带来复杂性。这种复杂性会极大地抑制用户参与构建知识库的热情,为了降低用户描述、发现和使用知识的难度,我们必须发展一系列辅助工具,利用这些工具来优化用户的应用体验。

为了发展图 1 方法体系所需的各类工具,我们首

先必须要解决一系列关键问题。其具体内容如下:

1)知识对象描述

在上述机构知识库中每一个知识单元都对应着一个或多个知识描述文档。这些知识描述文档主要描述知识单元的外表特征和内容特征。为了便于不同的计算机交换和处理知识描述文档,这些特征描述应该遵循一定的语法和语义规范。语法规则即是一种计算机可处理的数据表示方法和规则。语义规范则是与知识应用和知识管理过程密切相关的,它应该建立一套导入外部知识描述语义标签的方法。

2)知识标签组织

知识标签用以说明知识单元内容特征。由用户贡献知识标签带来了知识标签的丰富,但是也带来了严重的知识标签的无序问题。如何抽取用户知识标签以及组织知识标签就成为一个关键问题。为了便于计算机工具的处理,知识标签表述应该遵循一定语法规则;为了保障知识标签的语义理解一致性,知识标签的建立需要建立一定的管理程序;为了保证知识标签使用的便捷性,知识标签还必须要选择一种有效的组织方式。

3)基于知识描述文档的知识共享

知识描述文档建立了与其所描述知识单元之间的知识关联,这些知识关联为我们发现、应用和管理知识提供了有力的支持。知识描述文档的规范化以及其所用知识标签的组织性为构建知识检索和知识分类目录提供了有利的条件,但是应该看到,知识标签数据库中的半结构化的标签毕竟不是完全由一个主体建立的,其精确性要大打折扣。传统的基于标记的信息组织和共享方法不再适用,这就需要将知识对象描述规则和知识标签组织规则研究合适的知识共享方法。

3 KM@web2.0 --一种非中心化的机构知识库系统

3.1 KM@web2.0 系统概述

KM@web2.0 系统是我们承担国家社科项目《基于协同网络的行业知识库的构建及其应用研究》中设计的一个汽车行业知识库系统原型。该原型系统强调采用非中心化的策略借鉴机构知识库的结构来构建。

图 2 所示汽车行业知识库是由行业内各个企业知识库管理系统和汽车行业知识库管理系统通过协作共同组建而成。企业知识库系统中的知识由企业中的知

知识贡献者提供。知识贡献者在提交可共享知识单元的同时需要形成一份基于知识描述模式的知识描述 XML 文档。该描述文档是对知识特征的抽象和描述, 基于一定的知识描述元数据和规则。知识贡献者使用客户端工具将其知识描述文档和知识单元通过 SOAP 协议传递给其企业知识库管理系统。企业知识库管理

系统将知识单元存入企业知识库, 同时将知识描述文档上传至行业知识库管理系统。行业知识库管理系统将其存入知识描述库。行业知识库管理系统通过这样一种运作模式累积了行业内各个知识库中知识单元的描述文档集合, 通过这些知识描述文档可以直接链接到具体的知识单元。

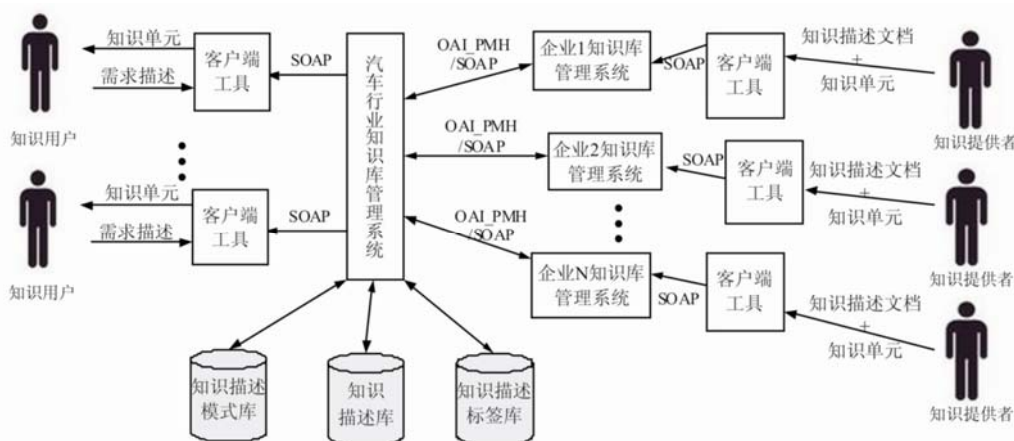


图 2 KM@web2.0 系统原型系统图

为了实现用户通过其客户端工具、行业知识库管理系统以及企业知识库管理系统这三者相互之间的灵活交互。我们使用了 SOA(面向服务的架构)技术, 即系统之间采用 SOAP 协议进行通信和交互。企业知识库管理系统将存入其数据库的知识单元首先使用 WSDL 协议将其封装成 web 知识服务, 然后通过开放接口将这样的 web 知识服务提供给广大用户使用。

为了增强知识管理和知识应用的效率, 我们的方法特别强调了知识的描述功能。知识贡献者在提交知识单元时按照知识描述规则和标签对其进行说明, 这样的结构化的知识描述就为我们使用各种分类、检索和挖掘技术手段提供了坚实的基础。行业知识库管理系统的主要功能在于通过其聚合的知识描述文档来构建行业知识地图, 给知识用户提供各种手段发现和使用其所需要的知识。

知识用户登陆行业知识库管理系统, 按照一定规则和标签描述其知识需求, 然后提交。这里我们强调了对知识需求描述的结构化处理。行业知识库搜索引擎按其到知识描述库去匹配相关的知识描述文档, 首先从所取得的知识描述文档中获取相应知识单元的网络链接 URI, 然后通过 SOAP 协议从具体的企业知识

库中获取该知识单元, 返回给知识用户。

3.2 KM@web2.0 系统的知识管理方法

1) 基于 XML 模式的知识描述文档

为了便于用工具实现对知识描述文档的半自动化和自动化处理, 知识描述文档应该是结构化的。我们选择 XML 作为表述语言, 借鉴 HTML 文档的结构特点, 建立了图 3 所示的知识对象标记文档结构图。

知识对象标记文档由标记文档头和标记文档体两部分组成。其中文档头部分主要用作工具处理, 文档体部分用于知识描述。标记文档头包含版本控制、管理特征、受控标签导入说明三部分。其中版本控制用于说明受控标签和标记文档的版本信息; 管理特征标记包括标记创建人、标记文档的权限, 标记的优先级等信息; 受控标签导入说明部分标记受控标签的来源, 这里借鉴 XML 的命名空间概念, 通过导入受控标签的 URI 来说明使用的受控标签。这里的受控标签主要应用于标记文档体部分的内容特征标记。

标记文档体分成形式特征描述和内容特征描述两部分。形式特征的内容借鉴都柏林核(DC)的数据元素, 并且对其做了简单调整。内容特征描述被划分为两个部分: 一为标准内容标记; 二为用户自定义内容标记。

标准内容标记主要有: 主题、摘要、关键词、所属分类、领域、上下文等, 这些内容特征标记是每一个知识对象描述文档所不可缺少的描述标记。



图 3 知识标记文档结构

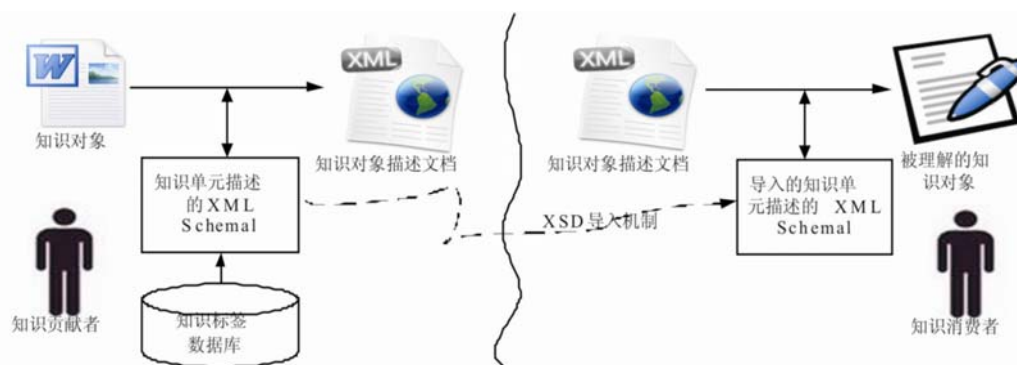


图 4 用户自定义标签导入机制

当知识消费者通过机构知识库发现了该知识对象, 他可以通过知识对象与其描述文档之间的关联关系发现该知识对象的描述文档。然后通过解读描述文档的文档头部分的受控标签导入机制找到解读该 XML 文档的 XML 模式文件的 URI, 通过网络将其下载到本机。最后利用客户端工具按照该 XML 模式来解读该知识对象的描述文档内容, 从而形成对原知识对象的理解。

2) 分层级的知识描述标签管理

借鉴社会化标签(Tag)的发展, 我们引入知识描述标签。不同于传统的知识描述元数据, 我们所导入的知识描述标签具有元数据的作用, 但是却强调通过社

用户自定义内容标记则提供了扩展描述知识的机制, 我们将用户自定义标记限定于所导入的受控标签中。内容特征描述部分的属性值也受到所导入的受控标签的约束。图 4 所示为用户自定义标签的导入机制示意图。

知识贡献者维护并管理自己的知识标签库。为了说明这些知识标签的使用规则, 他可以建立自己的用于知识单元描述的 XML 模式。每一知识贡献者可以按其需要建立若干不同的知识描述 XML 模式文件。当贡献者需要共享某一知识对象时, 它需要使用客户端工具描述该对象形成知识对象描述文档。在知识描述文档的文档头部分通过受控标签导入机制引入该描述文档所用的 XML 模式文件的 URI。并且保证该知识对象描述文档对该 XML 模式是有效的。知识贡献者将该知识对象递交机构知识库的知识存储库, 将该知识对象描述文档递交机构知识库的知识描述存储库。这两个库中的知识对象及其描述文档之间建立了一一对应的关联关系。

会化标注来生成。在 KM@web2.0 系统中采用的策略是将工具和人工管理相结合。

用户在使用客户端工具描述其知识单元时, 其所使用的描述标签没有任何的限制, 与社会化标注系统的应用没有任何差别。但是, 客户端工具自动采集这些标签, 并且将其与之对应的知识单元一起上交标签识别和管理系统。该系统对这些标记进行语义分析, 首先建立这些用户标签的自动分类。然后, 借助于统计分析工具以及人工管理来建立这些标签的语法和语义级别的关联性, 存入个人的知识标签数据库以备后用。

机构知识库中的知识描述标签分为三类,即个人标签、社区标签和全局标签。个人标签为每一用户所建立、维护和使用。使用它形成的知识描述文档需要共享者和消费者共享某一 XML 模式文件;社区标签为某一社区内大家达成共识的描述标签,在整个社区范围内共享。社区内成员可以直接使用该类标签;跨社区的用户间可以通过访问社区共享标签库来使用该类标签。全局标签是指在机构知识库内达成共识的知识描述标签,所有用户通过直接使用即可获得对其含义的共同理解。

用户所使用的描述标签可以由个人标签逐步发展演化成为机构知识库的全局标签。可以建立如图 5 所示的标签进化机制。社区标签来自于个人标签。首先,通过个人标签提交机制发现被社区内用户广泛使用的共同标签;然后,被提交的标签要通过社区标签表决机制作用以决定其是否升级为社区标签。按照同样的处理机制社区标签有可能演化成全局标签。这里最关键的问题是提交机制和表决机制的实现。提交机制的本质是常用标签发现算法,表决机制的本质在于建立一套可以使用机器推理实现的判断规则。

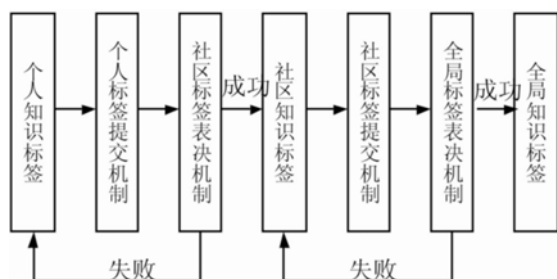


图 5 用户自定义标签进化机制

3.3 行业知识库应用服务

在 KM@web2.0 系统中知识单元存储以两种形式存在,一种是知识描述,其载体为知识描述数据库;一种是知识对象自身,其载体为知识存储数据库。知识描述数据库与知识存储数据库之间通过知识关联相联系。最简单的方法是采用超链接技术即可实现这种知识关联。这种知识关联的存在可以为我们发展更为丰富的知识服务奠定基础。

以图 3 所示原型系统为例,汽车行业知识库管理系统主要提供四大类知识应用服务:

1) 基于检索词的知识存储数据库检索。类似于传统搜索引擎,我们使用检索词的词频匹配作为返回检索结果的依据。

2) 基于受控标签的知识描述库检索。类似于中国期刊网的高级检索,根据我们所建立知识单元描述文档结构,我们引入多个检索域,并且使用知识标签数据库中的知识标签作为检索域可选项值,给用户提供一个精确描述其检索需求的端口。用户提交检索需求,检索服务器检索知识描述数据库返回满足其需要的知识描述文档,最后由用户决定是否需要导航至具体知识源。

3) 基于受控标签的知识分类目录。由于知识描述文档是结构化的,知识描述文档中的每一个元素都可以作为分类依据。此外,由于用户描述知识资源所用标签也是受控的,所以,所有知识标签数据库中的标签及其之间关系都可以作为组织知识对象的依据。为了适应这种多样性,我们根据知识描述文档模式设计了一种知识单元分类目录导航工具。该工具最大的特点在于它能够根据用户的不同应用需求,定制用户的分类目录导航需求。

4) 基于用户兴趣的知识推送服务。由于知识库中知识都是经过规范化和规则化约束所形成的,是可以描述的。用户可以根究自身需要按照知识描述规则描述其知识需求,将该需求递交服务器端可以实现知识的自动推送服务。推送服务得以实现的根源在于我们能够提供用户描述需求的工具,我们能够清晰定义用户的知识需求。用户描述其知识需求的工具是建立在前述知识描述文档模式的基础之上的。

3.4 KM@web2.0 系统服务器端的关键技术

KM@web2.0 系统的主体部分与中心化机构知识库没有本质的区别。只不过需要建立各种不同类型的知识库系统,并且要强调实现这些知识库系统之间以及其与客户端工具之间的灵活协作。

1) 基于标准的系统协作

为了实现各个知识库系统之间的协作,我们强调了标准的使用。SOAP、WSDL 是 SOA 架构下广泛使用的基于 www 网的远端程序访问和交互标准,它为实现 KM@web2.0 系统的软件层级的系统交互灵活性奠定了基础。OAI-PMH 协议是为获取特定元数据信息,有效挖掘和发布数字学术信息资源所定义的一个标准接口。OAI-PMH 协议的目标在于提供一种标准化的且简单易行的元数据互操作解决方案。该协议目前主要应

用于实现数字图书馆之间的互操作. 我们使用该协议来实现企业知识库系统和行业知识库系统之间知识描述文档的交互. 上述这些技术标准都有着广泛的技术资源可供参考, 故不再详述.

2) 基于 DSpace 的知识库实现

行业知识库和企业知识库仍然按照机构知识库的结构来构建. KM@web2.0 系统采用了开源软件 DSpace. DSpace 是美国麻省理工学院图书馆与美国惠普公司实验室共同开发的, 是迄今为止应用最为广泛的机构知识库构建的开源项目. DSpace 项目建立了机构知识库的资源框架、核心类库和对外接口, 用户所要做的是基于 DSpace 资源扩展自己的功能部分. 关于 DSpace 的技术资源相当广泛, 故不再赘述.

3.5 KM@web2.0 系统客户端工具的关键技术

KM@web2.0 系统中的客户端工具的关键功能在于帮助用户描述其知识对象. 纯手工的知识标记方法易于实现, 但是效率太低, 无疑会增加用户使用的难度, 这不利于本方法的推广应用. 为了降低用户使用的难度, 我们需要提供半自动化的知识标记辅助工具.

1) 基于 UIMA 的知识标记

客户端工具实现采用图 6 所示的基于自动分析的知识资源管理方法. 该方法的最大特征在于强调采用各种分析技术自动抽取知识对象的特征, 然后根据这些特征对知识对象进行自动标注, 将非结构的知识自动转换为结构化的知识描述, 在此基础上进行知识资源的管理和应用. 该方法完全依赖于各种自动分析机对知识对象的准确分析. 对知识文本进行分析需要应用自然语言处理、信息检索、智能文本分析以及知识挖掘等领域的各类技术, 而这些研究领域经过多年的发展已经积累了大量的技术成果可供我们使用. 现在的主要问题在于如何将现有的各类成熟的分析技术综合应用于我们的客户端工具中.

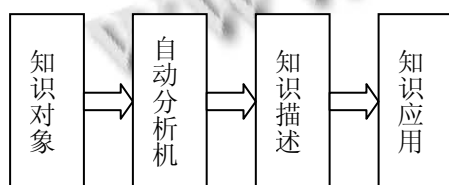


图 6 非结构信息资源管理的方法

在 KM@web2.0 系统中使用了 IBM 发布的开放源代码项目非结构信息管理架构 UIMA(Unstructured Information Management Architecture). UIMA 提供了在企业级的环境中处理各类非结构化的信息资源的通用

解决方法和支撑技术. Apache UIMA 提供支持 JAVA 和 C/C++ 的免费软件包下载^[11-13]. UIMA 框架是由 UIMA AE FACTORY 和各种 AE(Analysis Engines) 构成. UIMA AE FACTORY 负责各种 AE 的运行、调度和管理. AE 负责具体的数据源分析工作, 并且向用户提供标准的编程接口, 使用户可以编程访问分析的结果. AE 的核心是分析算法, 它承担分析文本和记录分析结果的所有工作. UIMA 提供一个基础的组件类型 type 来封装 AE 中运行的核心分析算法. 这个组件的实例被称为 Annotators. UIMA 框架提供合适的方法来使用 Annotator 和创建 AE. UIMA 提供的过程框架可以将各种简单的或复合的 AE 按照用户定义的分析过程集成为一个满足特定分析需求的复杂 AE, 这为 UIMA 提供了广阔的扩展空间. 用户使用 UIMA 的关键在于开发分析算法, 封装分析算法以及定义分析流程.

图 7 所示为基于 UIMA 设计的一种客户端工具系统体系结构. 该工具分为三大部分: 分析资源库、运行环境和组件库. 分析资源库中的资源组件以及组件库中的分析组件和应用组件都是通过标准接口集成于 UIMA 之上的.

2) 知识标签推荐

上述知识标记工具为用户提供了标记知识的支撑, 但是不能期望它能取代人的工作. 受制于当前的知识标记分析技术的不足, 在描述知识对象时我们还是要依赖人的手工标记. 如果能够提供标签自动推荐的功能就可以极大降低手工标记的难度, 此外, 在前述知识标签演化过程中所涉及的标签提交机制的实现也需要提供标签推荐的能力. 所以, 知识标签推荐模块就成为本系统的一个关键支撑工具.

标签推荐的本质在于根据现有的标记目标对可供选择的标签进行排序. 而对于标签的排序指标并不是唯一的, 可以考虑关键词, 词频, 领域本体, 以及词在文章中出现的位置信息等作为决策评价指标. 所以标签推荐问题可以看作是多目标决策问题. 在 KM@web2.0 系统中我们选择 TOPSIS 算法. TOPSIS 算法在众多领域得到成功的应用, 具有相当高的准确性和可操作性^[14].

4 结语

非中心化的机构知识库构建策略最大的优势在于其知识由用户贡献, 保证了知识库的知识多样性和丰富性. 但是, 用户自由贡献知识带来了知识库中知识的无序性, 这种无序性又严重影响了知识库中知识的应用. 论文提出的解决策略是: 由用户对其所贡献知

识按照受控标签进行描述和说明,实现知识的有序。为了解决知识描述标签建立的困难,论文借鉴社会化标签的思想,提出通过用户共享协作建立知识描述标签库,通过引入专门的工具和人工管理来实现这些描述标签的语义级别的确定性和关联性。论文对实现这一思想的 KM@web2.0 系统做了详细的介绍,以探讨其实现的可行性。通过我们对 KM@web2.0 原型系统

的研究,我们发现:非中心化的机构知识库要取得成功必须要吸引用户的广泛参与。尽管我们的方法降低了用户参与的难度,但这仅仅提供了用户参与的必要条件,还需要研究一套吸引用户参与的激励机制。未来我们将在非中心化的机构知识库的管理运作机理方面作进一步的探讨。

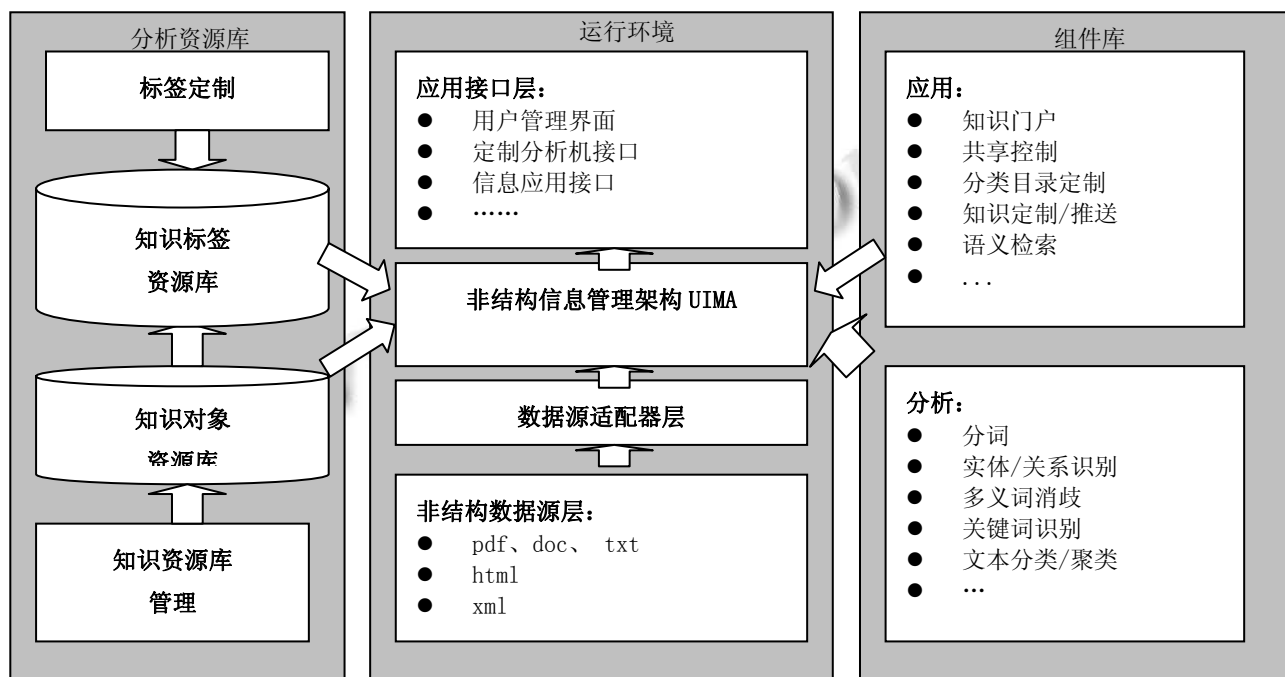


图 7 基于 UIMA 的客户端工具系统体系结构

参考文献

- Lynch CA. Institutional repositories: sential infrastrueture for scholarship in the digital age. ARL Bimonthly Report, 2003, (226): 1-7.
- 柯平,王颖洁.机构知识库的发展研究.图书馆论坛,2006, 6:101-104.
- 苏建华.机构库给高校图书馆带来的发展机遇.情报理论与 实践,2007,4:145-149.
- 侯壮.机构知识库与大学图书馆新拓展空间.现代情报, 2006,12:97-99.
- 洪梅,马建霞.机构知识库建设机制初探.情报杂志,2007, 8:32-35.
- 邓君.机构知识库(IR)系统 Archimede 与 eDoc 比较研究.图 书情报知识,2008,1:121-125.
- 蒋永福.知识秩序、知识共享、知识自由--关于图书馆精神 的制度维度思考.中国图书馆学报,2004(4):10-13.
- Cullen R, Chawner B. Institutional repositories, open access, and scholarly communication: A study of conflicting paradigms. The Journal of Academic Librarianship, 2011, 37(6): 460-470.
- Cocciolo A. Can Web 2.0 Enhance community participation in an institutional repository? The case of pocket knowledge at teachers college, columbia university. The Journal of Academic Librarianship, 2010, 36(4): 304-312.
- Xia JF. A ssesment of self-archiving in institutional repositories: Across disciplines. The Journal of Academic Librarianship, 2007, 33: 647-654.
- Ittoolbox Inc. Information Technology Toolbox. http:// knowledgemanagement.ittoolbox.com,2010.12.
- IBM. Unstructured Information Management Architecture (UIMA). http://domino.research.ibm.com/. 2007,12.
- Apache UIMA Development Community. UIMA Tutorial and Developers' Guides. http://incubator.apache.org/uima/ downloads/. 2007, 12.
- 赵宗仁.词语结构类比自动标引系统.情报学报, 1991,10(5):358-362.