

信息资源的组织与检索模式研究

张佩云^{1,2}, 吴江¹

(1. 西北大学 计算机科学系, 陕西 西安 710069; 2. 池州师范专科学校, 安徽 池州 247100)

摘 要:信息资源的激增给信息检索提出新的挑战。信息的检索离不开信息的组织形式。文中分析了信息组织的特点, 比较了几种信息检索模式的优缺点, 分析和研究了基于本体论的概念检索模式, 提出了该检索模式对应的信息组织和检索框架图, 并给出了框架图的一种实现方法。

关键词:概念检索; 本体论; OWL; 分类

中图分类号:G354.4; TP18

文献标识码:A

文章编号:1005-3751(2006)01-0132-03

Organization and Retrieval Model of Information Resources

ZHANG Pei-yun^{1,2}, WU Jiang¹

(1. Department of Computer Science, Northwest University, Xi'an 710069, China;

2. Chizhou Normal School, Chizhou 247100, China)

Abstract: The quick proliferation of information resource brings new challenge to information retrieval. Information retrieval is accompanying with organization of information. The paper analyses the features of information organization, compares advantages and disadvantages of information retrieval models, analyses and studies Ontology-based concept retrieval model, advances its framework of concept retrieval and offers an implementing method according to the framework.

Key words: concept retrieval; Ontology; OWL; hierarchy

0 引言

网络技术的发展使得网络信息的组织形式呈现多类型、多媒体、非规范、跨地区、分布分散、开放、无序等特点, 超媒体链接使得网络信息之间的关联性很强, 使得信息间呈现网状关联, 网络信息资源具有高度的动态性, 信息发布具有很大的随意性, 缺乏必要的过滤、质量控制和管理机制, 为用户选择利用网络信息带来了很大的不便, 使网络信息的查询、检索十分困难, 大大影响了信息利用的效率。如何更快更好地在 Internet 上查找所需信息是人们面临的一大难题, 因此只有对网络信息资源进行有效的组织管理, 才能实现信息资源效用的最大化。网络信息的存储多是数据库存储方式, 该信息组织方式带来数据结构不完善、重复记录多、规范性差、资源共享差等问题, 给检索带来不便。由于网络及其信息的产生、传播、管理的无序性, 使得用户难以从这种信息海洋之中获得特定信息, 给信息检索带来问题。信息检索时用户有时无法恰当、准确地表达出查询需求; 检索中一词多义和同义词的情况又无法正确检索到所需的信息。对于后者由于检索词本身与其概念的延伸不在同一级上, 使得利用传统信息检索所寻

找的信息造成字型的匹配, 但检索时想要的是检索项概念及相关的成分, 即传统信息检索关注“词”的处理而没有关心“词”的本原 (Ontology)。因此需要在知识的层面理解待检索词的含义, 需要采用新的信息的技术对信息进行表示、组织和分类, 信息组织是指采用一定的方式, 将某一方面的大量的、分散的、杂乱的信息经过整理、优化后形成的一个便于有效利用的系统的过程, 其目的是将无序信息组织成有序信息, 便于用户查找信息和有效地传递信息。

1 信息组织的方式以及文本信息检索模式分析

1.1 信息组织方式分析

网络信息资源进行组织使用得较多的方式主要有 4 种^[1]: 文件方式、数据库方式、主题树方式和超媒体方式。文件组织方式以文件方式组织网络信息资源简单方便, 可以降低信息组织的难度和成本, 能存储多种格式的非结构化信息。但是当随着网络信息资源信息量不断增多, 这种信息组织会使网络负载越来越大, 当信息结构复杂时, 就难以实现有效的控制和管理, 从而降低信息组织的效率。数据库组织方式是将所获得的信息资源按照固定的记录格式存储组织, 用户通过关键词及其匹配查询就可找到所需要的网络信息资源。该方式能高速处理大量结构化和非结构化数据。但是其面临重复性数据记录多、数据库结构不完善、数据库建设标准不统一、规范性差等缺点, 造成利用率低、资源共享差。主题树组织方式是从浏览界面机

收稿日期: 2005-04-12

基金项目: 陕西省教育厅产业化培育项目 (02JC47)

作者简介: 张佩云 (1974—), 女, 安徽安庆人, 硕士研究生, 讲师, 研究方向为智能信息处理、语义网。

制出发组织信息,用户通过该界面与网络信息资源的分类进行交流,并通过分类目录间接地连接并使用多个实际数据资源,分类目录组织成树型结构,用户按照规定的分类体系逐步查询,故查准率高,树型目录结构具有良好的可扩充性。超媒体组织方式是超文本技术与多媒体技术相结合的产物,它将多媒体信息以超文本方式组织起来,使人们可以通过高度链接的网络结构在各种信息库中自由航行,找到所需要的任何媒体的信息。

1.2 信息检索模式分析

信息检索(Information Retrieval),是指将信息按一定的方式组织和存储起来,并根据信息用户的需要找出有关的信息过程,这是广义的信息检索。狭义的信息检索仅指从信息集合中找出所需要的信息的过程,相当于人们通常所说的信息查寻(Information Search)。

目前信息检索主要模型有:布尔模型、向量空间模型、概率模型以及概念检索模型。

布尔模型是基于集合论的一种简单匹配模型,其缺点是无法在匹配结果集中进行相关性的排序,同时也无法区分词条在文档中所占的权重,并且漏检比较严重,可见布尔模型是一种简单但是不够理想的检索模型。向量空间模型中,文档用加权的关键词向量来表示,相似度用两个向量的夹角余弦来计算。该模型优点是比较简单,易于计算,但由于该模型术语间相互独立的前提假设有些过于简化,容易造成误检(检索到不相关的文档,例如在一词多义情况下)和漏检(没有检索到相关的文档,例如在同义词情况下)。概率模型是基于贝叶斯概率论原理的概率模型,不同于布尔和向量空间模型,它利用相关反馈的归纳学习方法,获取匹配函数。概念模型是采用网状结构来表示概念的组织和分类,搜索引擎根据该词语概念与其他词语概念的内在关联进行检索。使用概念模型检索,就不再局限于词条本身,当用户输入一个查询词条时,不仅要找出与查询表达式匹配的结果,也要找出包含与查询表达式概念相同或相近的词语的文档,即能实现语义检索。

从上述分析知:互联网上未组织的信息,无序程度非常大。从如此庞大的信息海洋中取出对用户最有用的信息,有必要使用概念检索。在概念检索过程中,不是采用字符匹配或相关的优化策略来查找目标,而是对检索对象进行语义处理,分析该语义段落中的潜在目标对象和查询请求的语义相关性,从而决定是否将其作为结果返回。

通过分析,将分类组织和概念检索相结合是提高信息资源重用和共享的有效方式。下面将针对目前分类体系存在的一些缺陷提出基于本体论的信息分类和概念检索模式。

2 本体论及元数据在概念检索中的作用

2.1 采用本体论思想对领域进行分类

2.1.1 现有分类体系存在的不足

为寻找特定的信息,需要事先对信息进行分类和抽

取,目前的分类体系存在不足,网络信息分类体系与传统分类法有一定的差别,传统分类法是以学科分类和逻辑划分为基础的严密而深细的分类体系,主要以印刷型文献为对象,是一维性的。网络信息是多维性的、交互性的、动态性的信息。网络信息分类体系是一种松散的、多维性的分类体系,缺少逻辑性和规范化,不能揭示信息之间的逻辑关系,它主要体现以下几个方面^[1]:

(1)类目设置缺少规律性。网络分类中,其类目设置往往同时采取多个标准,每个标准在使用时又并不完整,有时甚至列出不同等级的类目,使得同位类的设置显得很混乱,缺乏逻辑性和规律性,从而影响信息用户的查找信息的效率。

(2)类名不规范。有些类名有多个名称,其归属也很随意,很不利于信息用户的检索。

(3)类目分类没有注释或提示,不能直接找到所需类名,必须逐级翻寻等。

因此有必要建立一个规范的分体系,采用本体论的思想构造领域的分类将能可以克服现有分类的一些不足。

2.1.2 采用本体论思想对领域进行分类

为实现有效的信息管理,必须对信息资源进行分类,文中分类采用了本体论作为基本分类思想。Studer 对本体的定义是:本体是共享概念模型的明确形式化规范说明^[2]。该定义包含4层含义:概念化是通过抽象出客观世界中一些现象的相关概念而得到的模型,其含义独立于具体的环境状态;明确所使用的概念及使用这些概念的约束都有明确(显式)的定义;形式化是指计算机可读的;共享知识本体中体现的是共同认可的知识,反映的是相关领域中公认的概念集,它所针对的是团体而不是个体。本体的目标是捕获相关的领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇(术语)和词汇之间相互关系的明确定义。本体(Ontology)能够以一种显式、形式化的方式来表示语义,提高异构系统之间的互操作性,促进信息共享。

Ontology 是解决语义层次上信息和知识共享及交换的基础,具有良好的概念层次结构和对逻辑推理的支持,因而在信息检索,特别是在基于概念的检索中得到了广泛的应用。传统的全文检索技术基于关键词匹配进行检索,往往存在查不全、查不准、检索质量不高的现象,特别是在网络信息时代,利用关键词匹配很难满足人们检索的要求。但是建立了本体后,在知识层面或者说概念层面上建立相关领域知识的层次,以帮助用户获得最佳的检索效果,比如用户可以进一步缩小查询范围或扩大查询,获得上位信息、下位信息以及平级信息等。另外,带有语义的概念检索还包括歧义信息和检索处理,如“苹果”,究竟是指水果还是电脑品牌,将通过全文索引、用户检索上下文分析以及用户相关性反馈等技术来处理,高效、准确地反馈给用户最需要的信息。

总之,构造本体的目的都是为了实现某种程度的知识共享和重用。文献[2]认为本体的作用主要有以下两方面:

- * 本体的分析澄清了领域知识的结构,从而为知识表示打好基础。本体可以重用,从而避免重复的领域知识分析。

- * 统一的术语和概念使知识共享成为可能。

由上可见,Ontology 的核心作用在于定义了某一领域、领域之间的一系列的概念和它们之间的关系。在这一系列的概念的支持下,本体实现领域知识的分类,将提高信息搜索、信息资源的积累、信息共享。本体被用于知识表示,作为描述对象的语义载体,保证用户查询请求的语义完整性不受到破坏,同时借助本体对待查询对象进行判断。

2.2 元数据在概念检索中的作用

为组织互联网信息资源,围绕着 HTML 和 XML 环境产生了一系列元数据规范。其中较有影响的元数据标准如 Dublin Core(DC), PICS, WebCollections, CDF, 除此之外,还有许多应用于各种特殊领域的元数据规范。从网络环境下的信息组织与检索来看,将这些规范进行合理统一并制定出一种灵活的、能够支持多种元数据规范的标准,创造一个简单的元数据模型和体系方案显得非常必要。随着互联网上信息搜索服务的发展,在各种元数据格式和不同用户团体之间,也特别需要一种标准化的元数据集合或交换格式语言。DC 能较好地解决网络资源的发现、控制和管理问题,能被用来描述种种广泛主题学科和系统范围内的种种广泛主题,能由信息提供者或站点管理人员自己制作元数据。这对于数字图书馆的建设及其重要。都柏林核心的任一元素都是独立描述的,不依赖于具体的编码方法,与任何具体的传输结构都没有必然的联系。这样可以将 DC 映像转换为其它数据结构。DC 依据其所描述内容的类别和范围可分为 3 组,即对资源内容的描述、对知识产权的描述和对外部属性的描述。具体的 DC 元素有:题名(Title)、作者或创建者(Author Creator)、主题及关键词(Subject and Keywords)、描述(Description)、出版者(Publisher)、其它责任者(Other Contributors)、日期(Date)、类型(Resource Type)、格式(Resource Identifier)、来源(Source)、语种(Language)、关联(Relation)、覆盖范围(Coverage)、权限管理(Rights Management)。元数据在信息检索中扮演着重要角色,通过抽取信息,并用都柏林核心元数据表示,易于实现共享和重用,易于检索。

3 基于本体论及元数据的概念检索实现

基于本体论及元数据的信息组织和信息检索一般框架如图 1 所示。

语义 Web(Semantic Web)是万维网的发明人 Tim Berners-Lee 提出的新概念。按照 Lee 的描述,语义 Web 是对当前万维网的一个扩展,其中的信息都具有良好定义

的语义,这种语义是明确定义的,计算机能够理解并可操作的,显然这样的信息形式能够很容易地被检索和处理,提高人们使用网络的效率。为使得网页上的信息具有良好的语义,就要对网页上的信息进行语义描述。语义 Web 的本体描述语言为概念检索提供了领域共享概念的基础。语义 Web 体系结构是一个多层架构,文中利用其中的 3 层实现信息的组织和信息的检索。这 3 层分别是:标准信息交换格式层(XML)、信息表示层(如 RDF(S)^[3]、Topic Maps^[4]等)、本体层(如 OWL^[5]、DAML + OIL 等)。XML 是 Web 数据使用的通用语言,具有结构化、规范性、可扩展性及简洁的特点。它是在超级分布式系统之间实现多数据集传输的一种手段;它可同时使开发人员以更具价值的新型方式聚集和组合各种来源的数据。在 XML 的基础上采用 RDF(S)和 Topic Maps 描述资源元数据信息的元数据标准。由于信息资源需要很好的分类才能有效检索,因此采用本体层中 OWL、DAML + OIL 等本体描述语言来形式化的表示本体,实现本体库的建立。当用户检索时依照已经建立的本体和抽取的信息进行信息资源的概念检索以及推理检索。

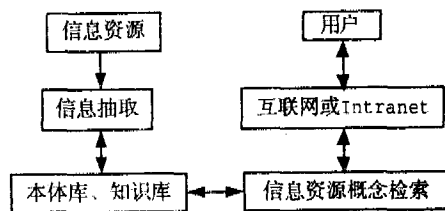


图 1 信息组织与信息检索框架图

以下给出信息组织与信息检索的框架图实现方法。

(1) 首先借助 OWL、DAML + OIL 等本体描述语言对应用领域进行形式化的、明确的描述,对应用领域资源信息的属性和联系进行定义,建立起该领域的本体模型,建立领域本体。建立本体时需要确定本体所覆盖的范围;考虑重用现有的本体;列举重要术语(概念),给出明确定义;明确概念和概念之间的关系(如基本关系 is-a, part-of 等);定义重要术语的属性和性质。构建本体时,Guarino^[6]等人提出了一套用于指导概念分类的可行理论。关于本体开发方法比较多,如 Michael Uschold 根据 Enterprise Ontology 的开发经验总结一个本体的开发方法就是其中的一种。

(2) 采用 XML 信息表示技术及通用的元数据描述工具 RDF(S)和 Topic Maps,对网上信息建立规范化表述。通过使用已经建立的领域本体以及采用元数据描述工具对网站信息描述后,使得网站上标注的或者未标注的信息具有语义和联系。如网站按照已经定义的本体对网站上的一篇文章进行规范化的描述,可对文献的外在特征如题目、作者、作者工作单位、专利和科技报告还有专利号或报告号等进行信息抽取,根据文献的内容特征对文献进行归类,如按照本体的类的层次对该文献所属领域进行归类

(下转第 157 页)

如图 6 所示。

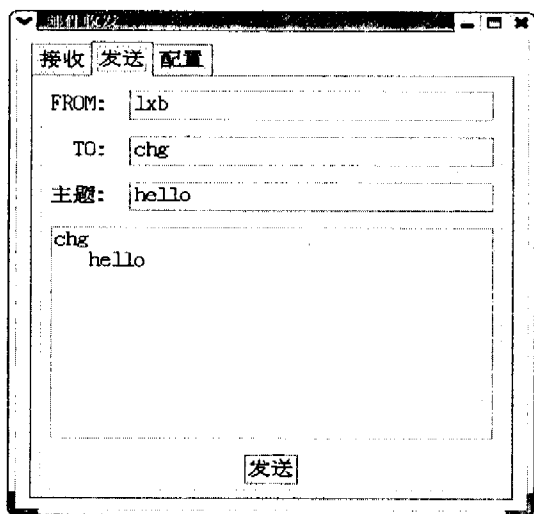


图 6 运行界面

基于 Linux 的邮件服务器其实有很多而且功能都非常强大,比如 Sendmail、Qmail 等,但是为了将其移植到嵌入式系统中需要做许多工作,其重点内容是研究邮件服务器的核心内容,如何在嵌入式系统中具体应用,而不是简单的二次开发。本系统的特点是基于 Linux 操作系统,兼具邮件服务器端和邮件客户端,另外邮件客户端还提供一个图形界面,方便了用户的操作。

本系统虽然在功能上实现了邮件的收发,但还存在着

(上接第 134 页)

等。通过利用元数据信息和分类信息将大大提高信息资源查找效率。

(3)信息检索实现。通过 Web 服务器端软件,为用户提供树型分类目录以及信息查询的界面。采用语义分析相匹配的方法实现概念检索,当输入自然语言进行检索时,分词处理是关键,主要是获取分词结果中的对象的语义,在查询请求和该对象在语义内容上的匹配和比较的过程中进行判断,检索过程将对待检索的对象进行语义分析,并与该对象所属分类的属性进行比较,得出该对象的判断结果并返回给用户。基于 Ontology 可以将同义词扩展检索、语义蕴涵、外延扩展检索、语义相关扩展检索等技术与信息检索结合起来,提供交互式的智能化信息检索服务,而信息检索的结果(如网页)可以作为知识检索的背景内容。

4 结束语

信息检索的分布化和网络化、开放性和集成性要求越来越高,使得要采用相应的信息组织方式和信息检索技术来检索和整合不同来源和结构的信息。信息检索在互联网信息日益增长的情况下起着非常重要的作用。文中研究了信息组织和信息检索技术,并针对部分存在的缺点,

一些缺点,诸如系统功能还不够完善,安全问题的解决、媒体文件的传输以及适用于嵌入式系统的后台数据库的研究,有待进一步完善。笔者将继续对本系统进行修改和完善,以实现在网络家电和智能化仪器仪表上的应用。

参考文献:

- [1] 刘文峰,李程远,李善平. 嵌入式 Linux 操作系统的研究[J]. 浙江大学学报(工学版), 2004, 38(4): 447-452.
- [2] 陈闯中. Linux 在嵌入式操作系统中的应用[J]. 同济大学学报(自然科学版), 2001, 29(5): 564-566.
- [3] Postel J B. RFC821[DB/OL]. <http://www.ietf.org>, 1982-08.
- [4] Myers J. RFC1939[DB/OL]. <http://www.ietf.org>, 1996-05.
- [5] Crocker D H. RFC822[DB/OL]. <http://www.ietf.org>, 1982-08.
- [6] Braden R. RFC1123[DB/OL]. <http://www.ietf.org>, 1989-10.
- [7] 于明俭,陈向阳,方 汉. Linux 程序设计权威指南[M]. 北京:机械工业出版社, 2001.
- [8] 周良源. UNIX 平台下的 C 编程指南[M]. 北京:电子工业出版社, 2000.
- [9] Wright P. GTK+ /GNOME 程序设计[M]. 钟 鸣,石永平译. 北京:机械工业出版社, 2002.

进行了改进。

参考文献:

- [1] 张 帆. 信息存储与检索[M]. 北京:高等教育出版社, 2003.
- [2] Studer R, Benjamins V R, Fensel D. Knowledge Engineering, Principles and Methods[J]. Data and Knowledge Engineering, 1998, 25(1-2): 161-197.
- [3] Lassila O, Swick R. Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation[EB/OL]. <http://www.w3.org/TR/REC-rdf-syntax>, 1999-02-22.
- [4] Park J, Hunting S. XML Topic Maps: creating and using Topic Maps for the web[M]. New York: Prentice, 2003.
- [5] Patel - Schneider P F, Hayes P. OWL[EB/OL]. <http://www.w3.org/TR/2003/CR-owl-semantic-20030818/>, 2003.
- [6] Guarino N, Welty C. Towards a methodology for ontology-based model engineering. In Proceedings of the ECOOP - 2000 Workshop on Model Engineering[EB/OL]. <http://www.ladseb.pd.cnr.it/infor/ontology/Papers/OntologyPapers.html>, 2000.