

基于 XML 技术的个性化主动信息服务系统关键技术的研究^{*}

陈俊杰^{1,2} 郭永明² 宋翰涛¹ 陆玉昌³ 余雪丽²

(北京理工大学 北京100081)¹ (太原理工大学 太原030024)² (清华大学 北京100084)³

The Study of Several Key Technologies of the Personal and Active Information Service System Based on XML Technology

CHEN Jun-Jie^{1,2} GUO Yun-Ming² SONG Han-Tao¹ LU Yu-Chang³ YU Xue-Li²

(Beijing Institute of Technology, Beijing 100081)¹

(Taiyuan University of Technology, Taiyuan 030024)² (Tsinghua University, Beijing 100084)³

Abstract With the rapid development and spread of the Internet technology, there have been significant progresses in the Internet information retrieval and active services. However, with the traditional Internet search engines and active services, which are all based on simple similarity technology of features and users' personal information are matching of keywords or bag of words, the document described through simple key words or words bag, and documents features and users' personal interests aren't well depicted due to the lack of semantics. XML is a newly emerged standard for the WWW data exchange and description. XML-based information retrieval system not only describes the contents of the documents, but also takes the documents' structural information. In this paper, a newer personal active information service system is discussed, under which the heterogeneous, distributed data spreading on the Web will be integrated to XML data, and users' personal information are described through XML queries. We have proposed the general architecture of this system, and several key technologies for the establishment of such a system are also presented.

Keywords XML, Personal and active information service, Description of users' interests

1 引言

随着网络技术的飞速发展,互联网的应用日益普及,网上信息呈现爆炸性增长的趋势,面对信息的汪洋大海,人们感到一片茫然,无所适从。为了能够帮助用户找到自己所需要的信息,各种各样的网上搜索工具应运而生。但是这些搜索工具往往返回大量的信息页面,用户要找到自己真正关心的信息还必须花费大量的时间来逐一浏览这些网页,有时候当你浏览完搜索工具返回的所有页面时却未必能找到自己真正需要的信息。信息检索犹如大海捞针,往往无功而返。造成这种现象的主要原因是,传统的信息检索工具缺乏语义。传统的信息检索常采用的检索模型,如向量空间模型、概率统计模型等,把文档看成是一堆相互独立的关键词的集合,忽略了文档固有的结构及其所代表的语义信息,仅依靠简单的关键词或关键词集的相似性匹配来检索文档。因此,传统的信息检索工具不能很好地描述和刻画文档信息与用户需求以及它们之间的相似性,使得用户被大量无关的信息所困扰而找不到自己真正关心的信息。

XML(eXtensible Markup Language)可扩展标记语言的出现,使得网上的信息检索工作充满了希望。XML作为互联网上标准的信息描述和交换语言,它不仅描述了文档的内容,还包含了文档的结构信息。XML文档的结构信息及其所固有的语义信息可用来实现更加准确和有针对性的信息检索^[1]。XML技术的使用,解决了文档和用户需求无法准确描述的问题,

使得信息检索可充分利用XML文档的结构信息来增加信息检索的语义。

2 系统总体结构

基于XML技术的个性化主动信息服务系统基本功能概括地讲就是将收集到的信息设法转换成统一的XML格式文档,然后按照用户的个性化需要对文档集进行过滤,从而将符合用户需求的信息主动地传送给用户。系统总体结构图如图1所示。

该系统主要由三部分组成:

(1)XML文档格式转换部分。此部分由XML文档转换器构成,它负责对收集到的网上各种异构数据源(各种数据库数据、HTML文档等)进行分析标记,进而将其转换成XML文档。

(2)XML文档信息过滤部分。此部分由XML文档分析器和XML文档信息过滤器组成。每一个新到来的XML文档,经XML文档分析器分析后,送至XML文档信息过滤器;XML文档信息过滤器快速读取用户资源文件,查找与该XML文档相匹配的用户,然后主动地将该XML文档推送给此用户。

(3)用户资源文件建立部分。此部分由用户资源文件建立器和用户资源文件分析器组成。用户资源文件是采用XML的查询语言XPath写成的查询,它用来描述用户的个性化信息需求。用户资源文件建立器通过一种数据结构模式的分类

^{*}基金项目:国家重点基础研究规划(973项目编号G1998030414)。陈俊杰 教授,博士研究生,主要研究方向:数据库与智能信息处理,分布式多媒体。

层次导航器来建立用户资源文件;用户资源文件分析器则是在对用户资源文件进行分析的基础上来建立索引,以便当新

的 XML 文档到来时能快速定位与该文档匹配的用户资源文件。

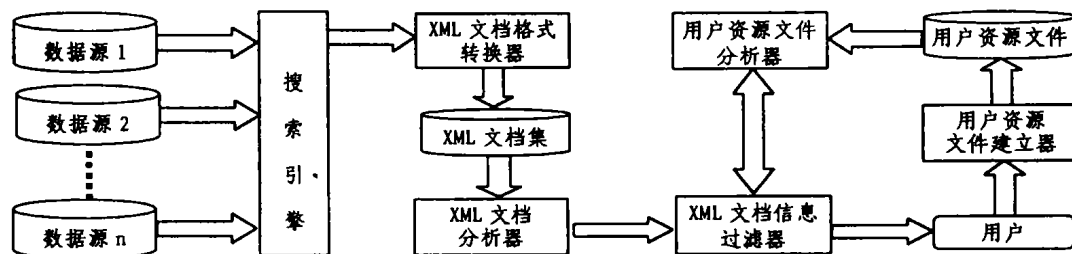


图1 基于XML技术的个性化主动信息服务系统结构框图

3 系统的关键技术及其实现

3.1 XML数据的转换

(1)HTML文档向XML文档的转换 HTML文档向XML文档转换时,一般先将HTML文档分析加工为DOM树的形式,在此基础上再作进一步的转换。转换过程主要是抽取HTML文档及其集合所表达的模式信息(也称为Web-Schema),这可通过程序自动从中抽取,抽取出来的信息可以比数据模式的粒度稍粗,但要比HTML文档及其之间的超链接的粒度细,其中有一部分是描述站点页面内容的信息,即各页面所要描述的对象及其一部分可分析出来的属性。例如,假设HTML页面每一段的第1个词都可以对应一个XML标记,有一部分是描述内容如何在HTML文档中组织的信息(如list,table等)。在此基础上,可以对HTML文本内容进行进一步的分析,将遗漏的模式信息补充进来,确定各对象的属性名和对象之间的语义关系,从而形成一个完整的数据模式。根据抽取出来的模式进一步确定结果XML的所有文档标记,指定HTML文档中待转换的信息所在的HTML标记,并明确这些被转换的HTML文档信息与XML文档标记的对应关系。根据这些关系,扫描HTML文档,并输出结果XML文档。

(2)关系数据向XML数据的转换 关系数据是有结构的数据,它形式简单,并可以在其上同时生成具有不同语义的信息视图,能很自然地用XML形式的数据加以表示,故其转换相对简单。但是由于关系数据不能直接表达语义,数据之间的语义由外在数据的完整性约束来定义,语义的实现需要通过关系数据的操作(如join连接操作)来完成,因此关键是将基于表和列的集合数据转换为含有丰富语义信息的XML层次嵌套结构的数据。SQL-92标准定义的外连接(outer-join)操作可以对这种转换提供支持。在分析数据库所包含的表间及表内列间关系的基础上,建立XML文档结构以及数据库结构和XML文档结构之间的对应关系。根据这些关系,扫描数据库,并输出结果XML文档,从而实现从关系数据到表达XML参照语义的XML数据的转换。

3.2 XML用户资源文件的建立

在本系统中用户资源文件是用来表示用户个性化需求信息的,它是由XML查询语言XPath写成的查询来描述的。

(1)采用XPath作为一种用户个性化信息的描述语言 XPath语言是用来对XML文档各个部分进行快速定位的一种查询语言^[2]。在本系统中采用XPath语言写成的查询来表示用户个性化信息需求,因此可以较为准确地描述用户的信息需求。

XML文档是一种树状的分层结构,XPath语言把XML

文档看成一个由节点组成的树,每个文档都有一个根节点(元素),它下面又包括子元素,元素可以嵌套至任何深度。元素和子元素都可以包含属性与文本数据。XML文档的结构由文档类型定义DTD(Document Type Definition)来描述,文档类型定义指定了元素名、属性名以及文本内容的特性。

XPath语言所描述的路径可以是根节点出发的绝对路径,也可以是从一个已知节点出发的相对路径。一个查询表达式由一系列的节点组成,如/网站/中文网站,就表示选择根节点下元素网站下的所有元素名称为中文网站的元素。

(2)用XPath语言写成的查询来描述用户资源文件 XML是基于分类体系的。每一个分类都有自己所特有的文档类型定义。系统通过提供一种结构模式的分类层次导航器,引导用户一级一级地找到自己所要查找的类别定义,使得用户可用此类别结构中的标签来定义查询。目前,用户资源文件的建立主要是通过系统和用户的交互来实现的,下一步将研究通过机器学习的方法实现基于Agent自主学习的用户个性化兴趣学习机制。

3.3 XML文档的过滤引擎

XML文档的过滤引擎是本系统的核心部分,使用了一种复杂的索引结构和改进的有限状态机(FSM)方法来快速定位和检查相关的用户资源文件^[3]。在本系统中先对用户资源文件进行分类和索引,然后形成包括所有元素名称的倒排索引,这个查询索引采用哈夫曼表来组织数据,用来对文档和当前的XPath查询进行匹配。过滤引擎把每一个XPath查询都转换成一个有限状态机的节点集合,在查询时,当一个有限状态机的最终状态达到时,就认为这个用户资源文件与文档匹配。限于篇幅,关于详细的过滤引擎的索引和匹配算法机制在此不再赘述。

4 相关工作

文[4]采用相关反馈学习算法和基于多用户个性化模式的层次智能信息滤波算法实现个性化信息过滤。采用用户与用户Agent以及用户Agent与信息Agent的交互机制,智能化适应用户兴趣的变化及环境的变化。

文[5]提出了一种基于三层Agent的个性化主动服务系统。利用智能Agent的自主学习功能对用户的信息需求进行分析,主动搜集用户平时感兴趣的信息,从中分析用户需求的个性化特点,建立用户个性化需求模式。

还有的系统采用基于机器学习的方法或人工智能的方法对用户的兴趣进行跟踪学习,以实现个性化信息检索。但所有这些系统在用户个性化信息和文档特征描述方面,都采用向量空间模型,把用户个性化信息或文档特征视为几个相互独立的关键词的集合,很难准确地描述文档和用户的个性化信

息需求,因此检索的准确性受到了影响,不能很好地满足用户的需求。

本文提出的基于 XML 技术的个性化主动信息服务模型可较准确地描述文档特征和用户个性化信息需求,从而能实现较为准确的信息服务。

结论 针对现在的大部分检索工具或系统都是基于简单的关键词匹配技术,本文提出了利用 XML 文档的结构信息及其所代表的丰富语义来改进传统的网络信息检索系统或工具的方法。在给出了系统总体结构的基础上,研究探讨了系统几个关键技术的实现方法。但是,由于网络信息的复杂性和 XML 文档结构的可扩展性等方面的原因,本文所研究探讨的

(上接第84页)

上述六个子层构成了电子商务的一种参考模型。每个子层专注于一个主题,不同电子商务参与者对不同子层负有不同的责任。比如,政府官员主要负责社会环境层和基础层;商业人士主要负责商业策略层;研发人员主要负责科学技术层、基本服务层和商业模式层。

3 语义实体

然而,成功的电子商务需要社会、经济、政治、科技等方面的通力协作^[2]。根据参考模型,这应该是六个子层的有机合成。因此,采用一种公共的方法描述和表示各个子层中的实例将有助于我们在各层间交换指标说明并有效合成一个电子商务平台。本文将这种概念定义为参考模型的语义实体。所谓语义实体(Ontology)是指“对某种共享概念显示的、机器可识别的规范说明”^[3]。

语义实体在 Web 工程中已被证明是一种必要的元素。它可以生成自然语言,集成智能信息,提供按照语义访问互联网的途径,并可以从文本信息中提取智能信息^[4]。语义实体的重要元素包括:概念(Concept)、关系(Relation)、函数(Function)、公理(Axioms)、实例(Instance)、事实(Fact)和声明(Claims)。

概念也被称为类、对象或分类,属性是其主要构成要素。关系反映了概念间的互操作性。函数是带有一个返回参数的特殊的关系。公理也被称为断言,用来约束信息、验证正确性、推导新信息。实例、事实和声明分别是概念、关系(函数)和公理的具体实现。

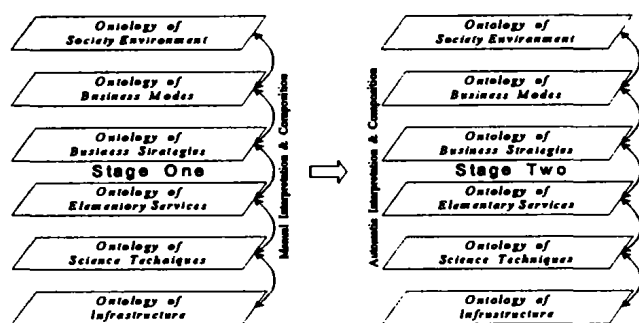


图2 使用语义实体的两个阶段

将语义实体引入参考模型后,可以分两个阶段从中受益。首先,需要将各子层实例用语义实体描述语言进行规范化,这样,各行各业可以方便地交流并有效地开发实现电子商务平台。

在第二个阶段,借助如图3所示的机制,可以在各层之间自动交换规范说明,自动合成电子商务解决方案。

几个关键技术的实现方法尚需做进一步的完善和改进,相信经过深入的研究,定将使其成为一个完善的系统。

参考文献

- 1 Extensible Markup Language (XML). <http://www.w3.org/XML/>
- 2 W3C (1999) "XML Path Language". <http://www.w3.org/TR/xpath>
- 3 altmel M, Franklin M J. Efficient Filtering of XML Documents for Selective Dissemination of Information. In: Proc. of the 26th VLDB Conference, Cairo, Egypt, 2000
- 4 汪晓岩,等. 面向 INTERNET 的个性化智能信息检索. 计算机研究与发展, 1999. 9
- 5 胡昌平,等. 基于三层 AGENT 的个性化主动服务系统的探索. 情报科学, 2001. 4

其中,语义实体注册器、编辑器和搜索引擎分别用来注册、编辑、查找特定的语义实体。所有的语义实体被保存在一个分层的知识库中,这个知识库可以具备分布式体系结构。如果需要,用户可以输入自己的商业轮廓文件,语义实体协调器将解读有关轮廓文件并自动合成满意的电子商务平台。

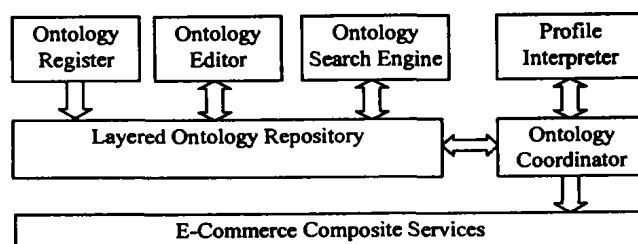


图3 通过语义实体实现电子商务自动化

为了实现上述图景,还需要为构造基本服务设计灵活的软件结构;用语义实体描述语言规范说明各个子层具体实例;设计语义实体、轮廓文件的查询匹配算法;解释不同层之间的语义实体;优化和仲裁同层语义实体之间的分歧和重叠。

结语 本文提出了电子商务的一个六层参考模型。尽管分类内容还不完善,但还是希望它有助于理解、设计、实现和参与电子商务。

将语义实体引入参考模型可以规范各行各业对于电子商务的描述并最终实现电子商务自动化。

借助上述参考模型,不难发现各行各业都可以从电子商务中受益。因此,电子商务是现代文明的一种必然趋势,并非“数字入侵者接管原本由其他国民经营着的商业领域的电子特洛伊木马”^[2]。

参考文献

- 1 Singh M P. An Evolutionary Look at E-Commerce. IEEE Internet Computing, 2001, 5(2): 6~7
- 2 Veeramani R, Talbert N. Where Are We in Global E-Commerce. IT Professional, 1999, 1(6): 46~52
- 3 Raman S. E-Commerce and Globalization-Yesterday, Today and Tomorrow. IEEE Engineering Management Society, 2000. 249 ~ 254
- 4 Menasci D A. A Reference Model for Designing an E-Commerce curriculum. IEEE Concurrency, 2000, 8(1): 82~85
- 5 Yang Jian, Papazoglou M P, van den Heuvel W. Tackling the Challenges of Service Composition in E-Marketplace. In: Proc. of IEEE 12th Intl. Workshop on Research Issues in Data Engineering: Engineering E-Commerce/e-Business Systems, 2002. 125~133
- 6 Wu Yingliang. Theory and Applications of E-Commerce. South China University of Technology Press, 2002
- 7 Siddiqi J, Akghar B, Davies C, et al. E-Commerce: Continuous Growth or Leveling Out. In: Proc. of IEEE Intl. Conf. on Information Technology, Coding and Computing, 2002. 491~496
- 8 Gomez-Perez A, Corcho O. Ontology Languages for the Semantic Web. IEEE Intelligent Systems, 2002, 17(1): 54~60