

基于 RDF 的个性化资源集成模型的研究与应用*

鲁 舟, 朱国进

(东华大学 计算机科学与技术学院, 上海 200051)

摘 要: 针对基于规则的个性化服务系统和基于信息过滤的个性化服务系统的若干缺点, 提出一种基于 RDF 的个性化资源集成模型, 为读者提供一个个性化、集成化的浏览空间, 同时给出了模型的一个应用, 该应用可以使互联网的用户拥有真正意义的“网上家园”。

关键词: 资源描述构架; 资源集成; 个性化; Agent

中图法分类号: TP393 文献标识码: A 文章编号: 1001-3695(2005)10-0069-02

Study and Application of a Personalized Resources Integration Model Based on RDF

LU Zhou, ZHU Guo-jin

(School of Computer Science & Technology, Donghua University, Shanghai 200051, China)

Abstract: Aiming at some disadvantages of the rules-based personalized services system and the information- flytering-based personalized services system, this paper put forward a model based on RDF for personalized resources integration, which proposes a personalized and integrated browsing space for reader and give a application of this model successfully. It makes the Internet users have theirs own " Network Home" .

Key words: Resource Description Framework; Resources Integration; Personalization; Agent

1 引言

随着 Internet 的迅猛发展, 网络用户正走向多极化, 网上资源越来越丰富, Web 成为人们获取信息和传播信息最重要的途径之一。由于 Web 信息的日益增长, 人们要找到自己所需要的信息, 必须花费大量的时间在浩如烟海的网络世界里搜索, 搜索引擎是最普遍的辅助人们检索信息的工具之一, 但由于其具有良好的通用性, 故无法满足不同背景、不同目的和不同时期的查询请求。个性化服务技术就是针对这个问题而提出的, 它为不同的用户提供不同的服务, 以满足不同的需求。个性化服务通过收集和分析用户信息来学习用户的兴趣和行为, 从而实现主动推荐的目的。个性化服务技术能充分提高站点的服务质量和访问效率, 从而吸引更多的访问者^[1~3]。

1.1 个性化信息服务的现状

现有的个性化服务系统根据其所采用的推荐技术分为两种^[1], 即基于规则的系统和信息过滤系统。信息过滤系统又可分为基于内容过滤的和协作过滤的系统。基于规则的系统如 WebSphere, BroadVision, ILOG 等, 它们允许系统管理员根据用户的静态特征和动态属性来制定规则, 规则决定了在不同情况下如何提供不同的服务, 基于规则的系统其优点是简单、直接, 缺点是规则质量很难保证, 而且不能动态更新。随着规则的数量增多, 系统将变得越来越难以管理。

基于内容的过滤系统如文献[4]中提到的 Personal Web-Watcher, 文献[5]中提到的 Syskill & Webert, 文献[6]中提到的

Letizia, 文献[7]中提到的 CiteSeer 等, 它们利用资源与用户兴趣的相似性来过滤信息。基于内容过滤的系统其优点是简单、有效, 缺点是难以区分资源内容的品质和风格, 而且不能为用户发现新的感兴趣的资源, 只能发现和用户已有兴趣相似的资源。

协作过滤系统如文献[8]中提到的 WebWatcher, 文献[9]中的 Let'sBrowse, 文献[10]中的 GroupLens, 文献[11]中的 Firefly, 文献[12]中的 SELECT 等, 它们利用用户之间的相似性来过滤信息。基于协作过滤系统的优点是能为用户发现新的感兴趣的信息, 缺点是存在两个很难解决的问题: 稀疏性, 即在系统使用初期, 由于系统资源还未获得足够多的评价, 系统很难利用这些评价来发现相似的用户; 可扩展性, 即随着系统用户和资源的增多, 系统的性能会越来越低。

1.2 RDF(Resource Description Framework)

RDF 是 W3C 于 1999 年颁布的一个 Internet 建议, 它提供了资源的通用描述方式^[7]。RDF 的中文全称是资源描述构架, 是用来描述资源及其之间关系的语言规范, 它不仅是描述数据的框架, 而且是表示数据的框架。元数据是“关于数据的数据”(Data about data)。它是相对于对象数据而言的, 是关于对象数据的一种概括、实质性的描述。MARC 数据是一种元数据。近年来, 随着 Internet 的发展和信息的丰富, 出现了许多元数据标准(格式)。如 DC, PICS 等, 这些元数据对网络资源进行描述、组织和整理, 使之有序, 方便利用。用户不必直接接触对象数据源就可以决定取舍。规范化的元数据对网络信息的组织、挖掘、检索和利用都十分有益^[13~16]。

鉴于各种元数据各自发展, 优势无综合利用, 并且内容有重复的状况, W3C 成立了 W3C Resource Description Framework

工作组, 负责研究并提出一个综合性构架来解决这个问题。因此, RDF 可以看作是一个元数据的容器。RDF 数据模型包括三个基本组成部分: 资源(Resource), 可以是一个完整或部分的网页, 也可以是它们的集合, 能通过 URL 引用的任何事物; 属性(Property), 描述资源的特性; 声明(Statements), 包括引用资源的指针以及该资源属性和属性值的表达式^[14, 15]。

面向 Agent 的计算能帮助人们在复杂、异构、不确定的信息环境中识别复杂模式。本文讨论如何使用 RDF 技术恰当的描述 Web 系统的信息, 使其在信息上能够被 Agent 理解, 在结构上能够被 Agent 解析和重组。

2 模型的研究与应用

2.1 模型的结构

本文讨论的基于 RDF 的个性化资源集成模型如图 1 所示。该模型实现了基于知识的分布式资源的集成, 比现有的个性化服务技术, 体现了较强的可扩展性与综合性, 同时可以实现动态的更新。

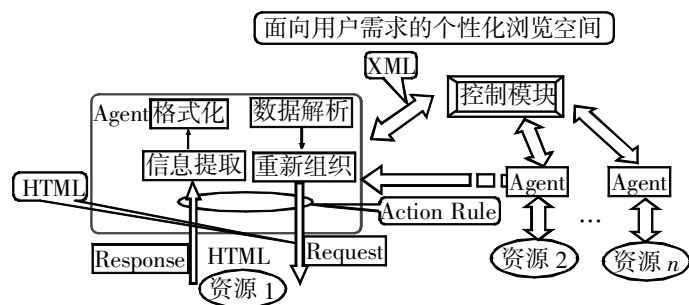


图 1 模型的结构

每一个被连接的分布式资源(即获取信息的 Web 站点)对应一个 Agent, Agent 通过由 RDF 定义的抽象接口与对应的 Web 站点进行 HTML 信息的交互, 而 Agent 内部则负责提取 HTML 信息中有用的部分并格式化存储到 XML 文档中。系统通过控制模块控制所有 Agent 的活动。

2.2 工作流程

(1) 资源描述。对资源(网页、网页中的一个元素或者段落)进行 RDF 描述。定义 RDF 文件如下所示:

```
< request url = " http: // news. sina. com. cn" >
  < rdf: BAG >
    < li > ... < /li >
    < li > ... < /li >
    < li > ... < /li >
    < li > ... < /li >
  < /rdf: BAG >
< /request >
< response url = " http: ... " >
  Status: Valladolid: IDSCMALN: 1000: - 1: 1. 00: 60: : " >
  < ex: parameter > userid < /ex: parameter >
  < rdf: BAG >
    < li > ... < /li >
    < li > ... < /li >
    < li > ... < /li >
    < li > ... < /li >
  < /rdf: BAG >
< /response >
```

(2) Agent 搜索与信息过滤。根据由 RDF 文件定义的 Agent 与 Web 系统交互的规则, Agent 从 Web 站点上相应的 URL 上得到 HTML 信息, 再从 Web 系统返回的 HTML 信息提取(Extract)有用的信息, 采用基于内容的信息过滤技术过滤掉冗余信息, 然后用 XML 的形式重新格式化(Format), 构造出通用

的结构化信息格式, 返回给控制模块。

(3) 信息映射。采用 Web 内容挖掘方法, 提取 Web 页中对应信息, 通过网络传送到本地以 XML 文档格式保存, 把 Web 页中相关信息映射到本地 XML 文档。

(4) 逻辑显示。自动整理本地数据库的信息, 按照读者的逻辑思维建立显示信息的 Web 页和 Web 页间的超链接。

(5) 自动更新。当控制模块收到 RDF 描述的更新请求后, 对请求进行解析, 读取其中第一个未处理的更新请求, 取得它的 Operation 属性。如果 Operation 是 Delete 则删除该资源的所有记录以及该资源包含的子资源记录(网页中的段落和元素); 如果 Operation 是 Update 则更正该资源的相关属性; 如果 Operation 是 Addnew 则新增该资源的记录^[17]。

从以上对模型的分析可以看出, 由于各个 Web 站点的异构性, Agent 与各分布式系统交互时交换的信息的结构与内容都是各不相同的, 但是经过 Agent 依照规则解析并重新格式化后, 与控制模块交互的 XML 信息的结构和内容都是一致的。这样的工作流程有利于模型进一步扩展, 可以将更多的 Web 资源分类聚集到该模型中来, 只要用 RDF 建立该系统的抽象模型, 即可利用 Agent 调用该资源。在维护上, 如果该系统的结构发生了变化, 只要修改与之对应的 RDF 文件, 即可使 Agent 适应系统新的结构。

2.3 模型的应用——建立用户真正的“网上家园”

对于互联网的大多数用户来讲, 都需要周期性地获取同类资源的最新信息, 如奥运会的金牌数量时时刻刻都会发生改变, 关于奥运会的新闻每天各个网站都要增加许多, 类似的还有天气预报等这些更新较快的资源。用户唯有及时手动地去搜索各个网页, 才能达到及时获取所需资源的目的, 即使这样做了, 结果还是往往不能令人满意。对于搜索的过程, 即使用搜索引擎, 一个一个手动去查看 URL 同样也是件令人烦心的事情。应用本文所提出基于 RDF 的个性化资源集成模型, 可以解决上述问题, 同时还能让您在网上建立一个自己的“网上家园”。

首先为用户建立一张 Web 页面, 页面上储存了用户想要得到的所有信息, 如奥运会金牌数、天气预报等; 然后通过 RDF 文件利用搜索引擎找到信息相关度高的网页, 即模型中的分布式资源, 根据模型的工作流程, 定期从该网页中获取用户想要了解的信息, 通过 Agent 将它送到本地, 以 XML 文件存储, 这样用户就可以在一张网页上实时了解所有想得到的最新信息了, 为用户提供了极大的方便, 使用户从大量、烦琐的搜索工作中解脱。当然对于该模型而言, 添加一个用户新的要求也是十分方便的。

3 结束语

本文提出了一个基于 RDF 的个性化资源集成模型, 它可以解决现有个性化服务存在的某些不足。为解决个性化的信息搜索、个性化的信息分类以及个性化信息的自动更新提供了很好的思路, 因而该模型将有较好的应用前景。

参考文献:

- [1] 曾春, 邢春晓, 周立柱. 个性化服务技术综述[J]. 软件学报, 2002, 13(10): 1952-1961.
- [2] Pretschner A. Ontology Based Personalized Search[D]. Lawrence, KS: University of Kansas, 1999. 123-128.

(下转第 73 页)

需要的是包含指向每个 Log2LogRule 类实例的指针列表。通过调用这些实例的 Log2LogRule 抽象类的虚方法, 搜索策略可以产生优化所需的所有逻辑操作符树。

3.2 物理实现规则

一个访问计划可以通过将逻辑操作符树中每个逻辑操作符实例替换为实现该操作符的一个物理操作符类实例而产生出来。替换工作由从 Log2PhysRule 抽象类派生的类来完成。

Log2PhysRule 抽象类同样包含虚方法 Apply() 和 CanBeApplied()。前者以一个逻辑操作符实例作为一个输入参数, 并创建一个或更多新的物理操作符实例, 它们代表使用物理执行算法来执行被逻辑操作符实例所代表的操作的不同的方法; 后者用来确定该 Log2PhysRule 是否可以被应用到给定的逻辑操作符实例上。

3.3 强制器规则

Phys2PhysRule 类在一个访问计划已被产生以后进一步修改它。其 Apply() 虚方法以一个物理操作符实例(代表一个访问计划) 为输入, 创建一个或多个新的物理操作符实例, 每个代表某个其他的访问计划。Phys2PhysRule 类的一个重要的作用是自动插入强制器实例, 它能够改变某些访问计划输出的物理属性, 以满足某些物理操作符的输入约束。例如, 在关系优化器中, 可以从 Phys2PhysRule 类派生一个类以强制在访问计划结果上的各种排序顺序。在优化过程中, 当搜索策略使用 Log2PhysRule 类来构建各种访问计划时, 只要一个新的物理操作符实例被创建, 搜索策略就先得得到其输入约束, 若输入不满足它的输入约束, 搜索策略试图通过应用一个适当的 Phys2PhysRule 来调整这种情况。搜索策略使用 Phys2PhysRule 类的 CanBeApplied() 虚方法来确定该强制器可否被用来强制给定的属性, 并且调用 Apply() 方法来创建新的满足对应的输入约束的访问计划。

4 总结

本文介绍了查询优化器的一个面向对象框架。它由代数子系统、查询重写子系统、搜索策略子系统组成。前两个子系统各自分为抽象层和具体层, 搜索策略子系统只使用抽象层中各种基类的虚方法来完成搜索最优计划的工作。我们给出了代数子系统和查询重写子系统的抽象层。以此框架为基础, 通过扩充代数子系统和查询重写子系统的具体层, 实现了一个查询优化器原型。实验表明, 使用面向对象的设计方法, 通过继承与多态机制可提供较好的可扩展性。不需要特殊的规则语言, 只扩展代数子系统和查询重写子系统, 就可以实现一个功能强大、性能优良的优化器。

参考文献:

[1] urajit Chaudhur. An Overview of Query Optimization in Relational Systems[C] . Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 1998. 34-43.

[2] Hamid Pirahesh, Joseph M Hellerstein, Waqar Hasan. Extensible/Rule Based Query Rewrite Optimization in Starburst[C] . Proceeding of the ACM SIGMOD International Conference on Management of Data, San Diego, Califomia, 1992. 39-48.

[3] Goetz Graefe, W J McKenna. The Volcano Optimizer Generator: Extensibility and Efficient Search[C] . Proceeding of the 12th International Conference on Data Engineering, 1993. 209-218.

[4] Lane B Warshaw, Daniel P Miranker. Rule-based Query Optimization, Revisited[C] . Proceedings of the 1999 ACM CIKM International Conference on Information and Knowledge Management, Kansas City, Missouri, USA, 1999. 267-275.

[5] Goetz Graefe. Query Evaluation Techniques for Large Databases [J] . ACM Computing Surveys, 1993, 25(2) : 74-170.

作者简介:

许向阳(1967-), 男, 副教授, 主要研究方向为现代数据库技术; 徐持恒(1980-), 男, 硕士研究生, 主要研究方向为数据库管理系统的设计与实现技术。

(上接第 70 页)

[3] 杨凯,滕至阳. 基于 Agent 的个性化信息服务[J] . 计算机工程与应用, 2002, (12) : 64-66.

[4] Mladenic D. Machine Learning for Better Web Browsing[C] . Rogers, S.Iba,W. eds. AAAI 2000 Spring Symposium Technical Reports on Adaptive User Interfaces. Menlo Park, CA: AAAI Press, 2000. 82-84.

[5] Pazzani M J Muramatsu, J Billsus, et al. Identifying Interesting Web Sites[C] . Weld, D. , Clancey, B eds. Proceedings of the 13th National Conference on Artificial Intelligence and 8th Innovative Applications of Artificial Intelligence Conference. Menlo Park, CA: AAAI Press, 1996. 54-61.

[6] Lieberman, H Letizia. An Agent that Assists Web Browsing[C] . Burke, R. ed. Proceedings of the International Joint Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 1995. 924-929.

[7] Bollacker, K D Lawrence, S Giles, C L. Discovering Relevant Scientific literature on the Web[J] . IEEE Intelligent Systems, 2000, 15(2) : 42-47.

[8] Joachims T, et al. WebWatcher: A Tour Guide for the World Wide Web[C] . Georgeff, M. P. Pollack, E. M. , eds. Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers, 1997. 770-777.

[9] Lieberman H, Dyke N V, Vivacqua A. Let 's Browse: A Collaborative Web Browsing Agent[C] . Proceedings of the International Conference on Intelligent User Interfaces. Los Angeles, CA: ACM Press, 1999. 65-68.

[10] Konstan J, et al. GroupLens: Applying Ollaborative Filtering to Usenet News[J] . Communications of the ACM, 1997, 40(3) : 77-87.

[11] Shardanand U, Maes P. Social Information Filtering: Algorithms for Automating Word of Mouth[C] . Roberts, T Robertson, S. eds. Proceedings of the ACM CHI 95 Conference on Human Factors in Computing Systems. New York: ACM Press, 1995. 210-217.

[12] Alton-Scheidl, R Ekhall, et al. SELECT: Social and Collaborative Filtering of Web Documents and News[C] . Proceedings of the 5th ERCIM Workshop on User Interfaces for All: User-Tailored Information Environments, 1999. 23-37.

[13] RDF/XML Syntax Specification (Revised) [EB/OL] . http: //www. w3. org/TR/ rdf-syntax- grammar/, 2004-09-08.

[14] Abstract Behavior Representations for Service Integration[EB/OL] . http: //www. agentcities. org/ EURTD/ Pubs/ eurtd. 02. constantinescu. abr. pdf, 2004-10-09.

[15] Richards, D Splunter, S van Brazier, et al. Composing Web Services Using an Agent Factory[EB/OL] . http: //www. agentus. com/ WSA-BE2003/ program/ richards. pdf, 2004-10-02.

[16] 姜恩波. RDF 原理、结构初探[J] . 现代图书情报技术, 2001, (5) : 32-33.

[17] 袁勇智. 一个基于 RDF/XML 自动更新的搜索引擎的设计与实现[J] . 农业图书情报学刊, 2003, (4) : 9-11.

作者简介:

鲁舟(1983-), 男, 辽宁人, 研究方向为计算机网络; 朱国进(1958-), 男, 上海人, 系主任, 副教授, 硕士, 研究方向为计算机网络。