

基于数字图书馆的本体应用环境研究 *

董 慧 王 菲 姜 赢 高 巾 翁丹丹 曾 杰

摘 要 语义 Web 框架虽已提出多年,但大部分应用都只是系统模型,很少投入实际的商业运作。为此,笔者从实用的角度提出本体应用环境的概念与框架。本体应用环境以本体技术为核心,以知识管理为目标,较好地解决了各领域知识管理中的各种问题,如知识组织、知识检索和知识推理等。目前,本体应用环境已经成功应用于数字图书馆、数字档案馆等领域,为语义 Web 的发展提供了一种可行的思路。图 4。参考文献 15。

关键词 本体应用环境 本体 本体分子 知识管理 语义 Web

分类号 G250. 76

ABSTRACT This paper analyzes the absence of Killing Application in Semantic Web, and advances a theory as well as a framework named Ontology Application Environment (OAE) from the aspect of real applications. Aiming at knowledge management, OAE makes ontology as the key technology. It resolves such knowledge management problems in many application domains as knowledge organization, knowledge retrieval, knowledge inference, etc. This paper describes OAE in detail, from its construction, back-end, foregrounding and web environment; and also puts forward Ontology molecule, which suits for static and dynamic knowledge management. OAE has been applied successfully in the domain of Digital Library and Digital Archives, which provides a feasible approach for the development and the future of Semantic Web. 4 figs. 15 refs.

KEY WORDS Ontology application environment. Ontology. Ontology molecule. Knowledge management. Semantic web.

CLASS NUMBER G250. 76

1 研究背景

语义 Web 的概念由 Tim Berners-Lee 在 1998 年首次提出^[1]。他认为“语义网 (Semantic Web)并非另外一个是独立的 Web,而是现在的 Web 的延伸。”2000 年秋,国际万维网联盟 (W3C)在美国波士顿召开会员大会。在此届年会上,W3C 总裁 Tim Berners-Lee 宣布语义网是 W3C 的三大研究主题之一^[2]。随后 W3C 制定了语义 Web 体系结构标准,它是一个层次结构 (Layer Cake),如图 1 所示,自底向上主要包括:

① URI/IRI:国际编码提供世界各种语言和符号的编码库,统一资源标识保证网上资源的

唯一性。这两个不仅仅是语义 Web,而且已经是当前 Web 实现的基础技术保证。

② XML/XML Schema 解决数据类型并定义数据结构的问题 (Datatype and structure)。

③ RDF/RDFS:RDF 利用三元组的方式定义和描述网络资源和元数据。RDFS 在 RDF 基础上为其提供基本词表,它是一种原始的本体定义框架。

④ OWL,Ontology:OWL 以描述逻辑作为数学逻辑基础,并在 RDFS 基础之上,提供更加强大复杂的基本词表,比如关于类、属性、关系的约束^[3]。

⑤ Unifying Logic:提供逻辑推理,即对于没有显式定义的知识进行推理。

* 本文系国家自然科学基金资助项目“基于数字图书馆的本体演化和知识管理研究”(批准号:70773087)成果之一。

⑥ Proof:提供逻辑验证。它跟踪逻辑推理过程,对逻辑推理结果的正确性进行验证。

⑦ Trust, Digital Signature:对下层所有的内容进行数字签名,使得用户信任语义 Web 处理的结果和质量。

⑧ User Interface:应用级别的用户接口,或者与领域相关的应用程序。

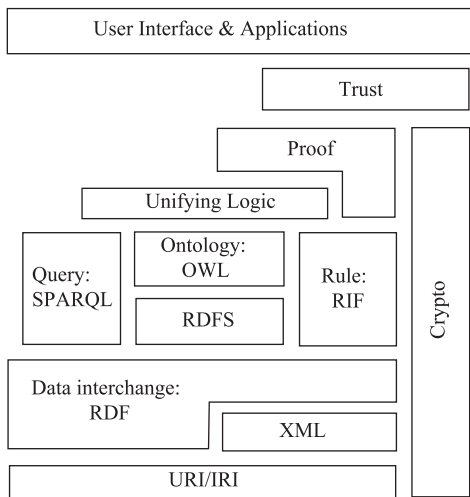


图1 语义 Web 框架

1999 年,W3C 的研究小组提出了资源描述框架(Resource Description Framework,RDF)标准草案^[4]。2001 年,W3C 又开始着手制定 OWL (OWL Web Ontology Language)标准。2004 年 2 月 10 日,W3C 万维网联盟发布了 RDF 推荐标准和 OWL 推荐标准^[5]。本体和语义 Web 的关系可以描述为:本体是语义 Web 体系结构中的一个层,语义 Web 利用本体解决 Ontology 层的语义问题。我们通常所说的本体层,指的是 OWL 层,而 OWL 和 RDF/RDFS 都能在一定程度上表达语义,所以也可以说 RDF 和 RDFS 属于本体层。

语义 Web 的框架已经提出若干年了,但是目前其他领域却鲜有基于语义 Web 的典型应用^[6]。大部分应用都只是系统模型,很少投入实际的商业运作。例如 Foaf 这类的模型能让人们感觉语义 Web 的新鲜,但是真正的语义 Web

离我们还比较遥远^[7]。我们认为,语义 Web 框架在实际应用中存在一些问题:① 语义 Web 框架是一个理论模型,它趋于完善,但仍侧重于理论框架结构而非应用。② 语义 Web 框架中有一些层次,如 proof 层和 trust 层,很少在实际中用到,不利于语义 Web 框架的实际应用。③ 语义 Web 框架侧重于对现有 Web 的扩展,而对于非 Web 应用内容的知识管理涉及较少。④ 语义 Web 框架对于知识抽取、知识可视化、动态知识管理等实际问题没有提出相应的解决方案。

为此,我们提出本体应用环境的理论与框架,并在数字图书馆和数字档案馆等领域得到了应用。实践证明,本体应用环境的理论与框架具有可行性、可操作性和科学性。

2 本体应用环境

本体应用环境以本体技术为核心,以知识管理为目标,较好地解决了各领域知识管理中的各种问题,如知识组织、知识检索、知识管理和知识推理等。其框架如图 2 所示。

2.1 框架结构

图 2 所描述的本体应用环境不同于语义 Web 框架单向的层次结构,它分为前台和后台两种应用环境。后台环境为后台管理员提供本体库等的建立、维护工作。前台环境为用户提供检索策略和手段,以及可视化显示;为了实现多领域知识检索,需要提供领域知识分析器。

2.2 后台环境

后台环境是为管理员提供本体库的建立和维护、领域规则库的建立和维护、领域知识抽取等功能。我们采用 Protégé 开发组件作为各功能模块的插件^[8]。

2.2.1 数据源

本体应用环境支持的数据源可以是自由文本、XML、数据库、TIFF 图片、PDF 文档、Word 文档、Web 网页等。

2.2.2 知识库与规则库

为了解决动态知识的管理和控制,我们提

出了本体分子理论。本体分子是指在基本元素(本体实例、三元组)的基础上,根据语义或者语用划分,用唯一标识符标注的、无缺失的、最小冗余的本体知识单元。本体分子是基本元素和本体库之间的一个平衡点,它使得动态知识管理成为可能。一个本体分子由两个部分组成:核子和离子。本体分子通过核子来描述不变部分,通过离子来描述可变部分,能够较好地展现领域中的动态知识的演化过程和演化结果。一个本体分子可以没有离子,但是必须有且仅有一个核子。

规则库是存储本体推理规则的地方。推理规则是针对特定领域而制定的、描述领域隐性知识的规则。在本体应用环境中,我们用 Jena 自定义推理规则实现。

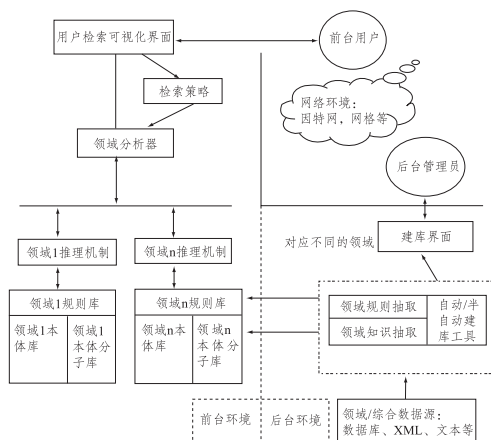


图2 本体应用环境框架

2.2.3 半自动建库工具

我们开发了基于 XML 模式匹配的中文本体半自动构建系统——OntoLTCn^[9],它是本体应用环境中进行领域知识工具抽取的工具^[10]。

OntoLTCn 运用了中文词法分析技术,针对中文语言特点,利用统计学的方法,建立了适合于中文文本的 XML 模式的匹配规则。插件从上下至下分为三个逻辑层:数据预处理层、本体抽取层和领域专家人工干预层。下层的处理步骤要依赖上层的处理结果,半自动建库过程由上向下依次进行^[11]。

2.3 前台环境

前台环境是为用户提供服务的窗口。我们设计的环境符合 J2EE 标准的 MVC 体系结构,以及 B/S 模式的 Web 开发,也符合 W3C 制定的 Web 设计标准。用户能在前台环境下通过友好的界面方便快捷地进行知识检索^[12]。

2.3.1 检索可视化

检索可视化是为用户提供丰富多样、形象生动的过程结果描述,使用户操作更加直观、便捷。

检索可视化采用了 Touchgraph 的 TG 图形组件^[13]。它能显示大量数据所需要的技术:缩放(Zoom)与摇动(Pan),焦点/背景(Focus/Context)及递增性导航(Incremental Navigation)等。

本体分子可视化的目标是提供领域动态知识演变过程的可视化。我们采用 Prefuse API^[14]。为解决不同对象的转换问题,可视化模型在 Prefuse 的基础上添加了数据对象映射层,负责本体和本体分子对象与 Prefuse 数据对象的互相映射转换。数据模型层沿用 Prefuse 本身的内容。抽象可视化数据模型层进行了二次开发,加入了本系统特有的颜色、字体、动画等抽象数据类。

2.3.2 检索策略

检索策略(retrieval strategy)是检索步骤的科学安排。它是为实现检索目标而制订的全盘计划或方案,指导整个检索过程。本体应用环境在前台环境中提供多种检索途径,用户可以从中选择合适的检索策略进行检索。

2.3.3 领域分析器

本体一般解决的是某个领域问题,它通过描述领域内部知识进行知识组织和知识检索。但一般用户在使用基于本体的综合领域检索系统时,往往并不知道自己要检索的内容属于哪个具体领域,或是否是跨领域知识。对于这些问题,本体应用环境应该提供领域分析器的解决方案。

领域分析器的实现可以分为几个级别层次:

(1) 初级实现:系统提供一个领域列表,用户可以根据自己的理解选择特定的一个或若干

个领域进行检索,检索的结果是这些领域所得结果的并集。

(2) 中级实现:系统后台有一个领域词表,它是领域和领域叙词的映射表。用户输入的关键词通过这个词表进行匹配,找到相应的领域之后,领域分析器再把检索请求分发到这个领域进行检索。

(3) 高级实现:系统后台配备一个语义智能分析器,可以根据用户输入判断它所属的领域,甚至多个领域,然后把请求发送到相关领域进行分别检索,最后将检索结果返回给前台用户。

领域分析器的实现层次可以根据应用的需要进行选择,或者采取逐步实现的方式。

2.3.4 本体推理

推理是按某种策略由已知判断推理出另一种判断的过程。本体知识推理是展示领域知识中所蕴涵的大量事实和复杂关系,对本体库进行维护、更新、检测和知识发现的一个十分重要的技术手段。

本体推理主要包括基于描述逻辑的推理和基于规则的推理。

我们采用了基于 TABLEAU 和 RETE 算法的推理机制。其主要功能有:本体推理引擎、本体逻辑检测分析器、本体推理路径回溯分析器。

2.4 网络环境

网络环境是整个本体应用环境所需的网络基础条件,还包括用户和管理员在本体应用环境下进行实际操作的条件。本体应用环境所支持的网络环境可以是因特网、局域网甚至网格环境^[15]。

3 本体应用环境的应用

本体应用环境是一个完整且可组装的架构,它能够有效指导基于本体的应用系统的设计与实现,而且可根据应用环境的具体领域,进行灵活组装。例如,只有静态知识的系统,本体分子库就是本体库,和开发一般的本体应用系统一样;对于既有静态知识又有动态知识的系

统,就需要加入扩展部分的框架。下面介绍的两个实例正是这两种应用的代表。

3.1 基于本体的数字图书馆信息检索模型研究

“基于本体的数字图书馆信息检索模型研究”,是国家自然科学基金资助项目(批准号:70373047)。该项目研究以历史领域(国共两党第二次合作时期)为突破口,将本体应用环境应用于数字图书馆领域,解决了对人物、事件等复杂关系的揭示与检索问题;通过推理机制丰富了本体库的内涵,完善了知识发现、组织、管理功能,提高了知识检索深度;而且还提供静态和动态的可视化方式,使用户操作更加直观、便捷。

该研究从知识管理角度看,由于历史资源是一个不变的知识集合,所以它要解决的只是静态知识的管理问题,因此,系统的本体应用环境框架,是一个无本体分子的静态知识管理模型,如图3所示。

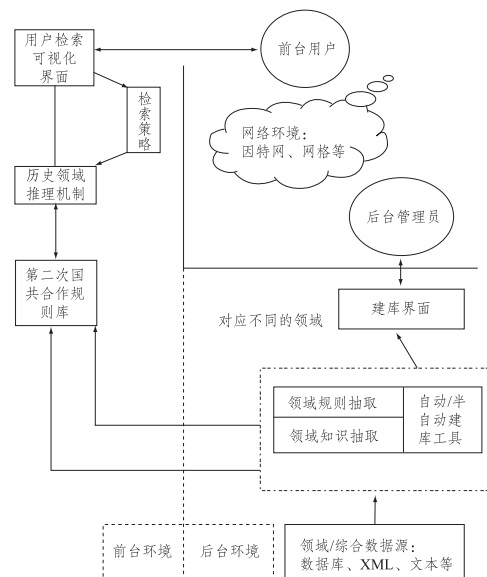


图3 基于领域静态知识管理的本体应用框架

后台环境中的数据来自《国共合作通史五卷本》和《中国革命历史文献采集标准》等,近百万余字。这些数据的特点是:

(1) 国共合作领域知识属于人文历史学科, 这个学科内的知识大部分都是已经发生过的静态知识。

(2) 国共合作历史时期时间跨度长, 有 90 年之久, 内容繁多, 史料浩瀚, 为了集中表现合作特点, 我们选择了第二次合作期间的人物和事件。

(3) 国共两党关系纷繁复杂, 牵扯甚广。

为了描述这段特定历史阶段的知识, 项目采取 OWL 本体技术构建了本体知识库, 并通过 Jena 自定规则技术构建了推理规则库。

本体库包括了 167 个本体类、108 个关系属性、100 个推理属性和 13838 个实例。其中, 本体类中有六个核心类: 事件 (Event)、资源 (Resource)、组织 (OrgActor)、人物 (IndiviActor)、时间 (Time)、地点 (Location); 实例中有事件本体实例 769 个, 资源本体实例 683 个, 组织本体实例 984 个, 人物本体实例 2137 个, 地点本体实例 1026 个, 时间本体实例 3645 个, 角色类实例 2838 个, 其他类实例 4567 个。平均关系复杂度为 5。推理规则库包含静态规则 102 条, 动态规则 5 条, 其中人物关系推理规则 54 条, 地理方位关系推理规则 15 条, 事件发展演变关系推理规则 14 条, 资源关系推理规则 4 条, 与时间有关推理规则 16 条, 组织演变规则 4 条。

在知识推理方面, 根据本体的推理机制可在基于逻辑的领域知识检测推理和基于关系的蕴涵知识发现推理这两类中进行选择。基于逻辑的领域知识检测推理, 主要是在 OWL 语言的逻辑基础 (描述逻辑) 的基础上, 运用 TABLEAU 算法, 对我们前期所构建的国共关系领域的概念层次、声明的实例以及实例间复杂的语义关系进行检测, 以保证本体库结构的逻辑一致性和知识描述的正确性, 为后续的蕴涵知识推理、本体库的拓展以及与其他领域本体库的集成提供逻辑保证。

基于关系的蕴涵知识发现推理的主要思想, 就是在国共合作领域本体库的概念层次与属性关系的基础上, 采取规则推理策略, 在本体库中进行模式匹配。在推理过程中, 主要考虑概念与概念之间关系 (超类、子类、成员、部分等

关系)、概念与个体关系 (实例关系)、实例与实例的关系 (亲缘、同事、相关、包含等关系) 以及对象属性关系 (父子、传递、互逆、对称等关系)。在蕴涵知识发现推理中, 可以判断个体是否是某个或多个类的实例, 确认某个类中所有的实例、推断两个实例之间的关系, 判断与某个实例有特定关系的实例、类 (属性) 的层次体系结构推理等, 根据本项目的特点和领域专家的建议, 我们选择了基于关系的蕴涵知识发现推理, 并建立了 107 条推理规则。

前台环境的检索入口主要包括: 列表检索、本体检索、关系检索 (可选择关系的深度为 1 到 5)、推理检索、回溯检索等。

由于本项目是针对历史这个特定领域, 不需要采用领域分析器将检索请求分配到各个领域。

本体推理检索是本系统最有特色的一种检索方式, 这种知识层次的推理是以往任何一种信息层次的数据库系统所无法做到的。系统除了提供本体推理检索的结果之外, 还对本体推理路径进行回溯, 即利用文本和可视化相结合的方式展现推理过程。另外, 在展开推理关系之后, 在系统中还显示出“推理路径回溯”的相关文字说明, 主要包括“推理规则”和“推理依据”。

3.2 基于数字图书馆的本体演化和知识管理研究

基于数字图书馆的本体演化和知识管理研究 (批准号: 70773087) 是国家自然科学基金资助项目, 该项目以数字图书馆在电子政务中的应用为背景, 研究电子公文应用过程中的知识管理及其演变。该项目与前一个项目相比是不同的, 其对不同类型知识进行管理, 且知识管理的内容不仅有静态知识, 还包括动态知识, 即需采用本体分子理论开展知识管理, 其本体应用环境框架如图 4 所示。

本项目关于静态知识的采集、组织、存储等与一般本体库的建设一样, 所不同的是关于动态知识采集、组织和存储的方法及措施。本项目提出了“知识组织四层模型 MRID”, 即元数据

层、知识表示层、推理层和动态知识层。动态知识主要由动态三元组、三元组标识符和三元组维度构成。根据本体分子理论开发了一个本体分子建库的工具 OMProtegePlugin,该工具具有:本体分子持久化,本体分子建库工程,动态三元组,核子和离子和本体分子等管理功能。这些功能较好地完成了对动态知识的管理,并建立了动态与静态知识间的互动机制。以上研究内容在本系列论文Ⅱ—动态知识组织中,进行了详细论述,在此不再赘述。目前,本项目的电子政务公文本体库包含了25个本体类、17个关系属性和26128个实例,案例中包括文件类11887个、条目类13821个、问题类7个、组织机构类402个、问题反映类11个;电子政务公文本体分子库包含了45个本体分子、25个核子和45个离子,核子中的三元组个数平均约为20个,离子中的三元组个数平均约为50个。

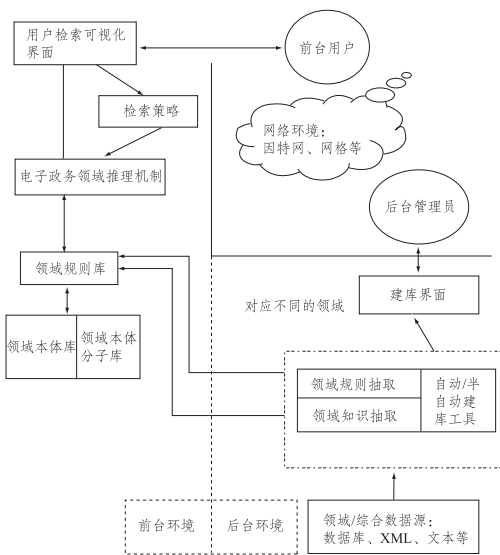


图4 基于一个领域动态知识管理的本体应用框架

数字图书馆在电子政务中的应用,其知识管理必须遵循该领域的规则。规则是制定推理机制的核心。例如,中央下发的文件(本项目研究的文件属非密级文件)省级政府是无权进行任何修改的,省级制定的文件在执行中可进行

修改、补充,甚至对整个文件或某些条文进行更正等。根据规则,中央文件和省级政府贯彻中央文件精神可构成一个本体分子。

对于前台检索策略,用户可以选择“一般检索”和“高级检索”两种方式检索文件。一般检索是采用关键词匹配文件名,高级检索则进行更多的元数据匹配(如发布者、检索号、责任人等)。用户检索到文件之后,还可以选择“文件关系检索”、“条目检索”、“条目关系检索”、“文件阶段演变检索”和“条目演变检索”等。其中,“文件关系检索”、“条目检索”、“条目关系检索”等都是针对知识库中静态知识的检索,而后两种针对动态知识的检索。

“文件阶段演变检索”是指对于本体分子描述的文件动态的变化过程的检索。例如,对于“再就业”问题,中央以及湖北省制定了相关政策和规定,它包括三个演化阶段:

- 2003-05-12 :中央制定了《国务院办公厅关于加快推进再就业工作的通知》。
- 2003-11-18 :湖北省在“再就业”这个问题上,贯彻了中央制定的《国务院办公厅关于加快推进再就业工作的通知》精神,结合本地特殊情况制定了《省工商局关于执行再就业扶持政策有关问题的通知》。
- 2004-04-09 :湖北省由于某些原因,制定了《省工商局关于执行再就业扶持政策有关问题的补充通知》,对上一个文件作了补充。

用户通过点击选择时间限制,进而检索出关于“再就业”问题各个不同时间阶段电子公文的演化过程。

针对电子公文会随时间变化在原来基础上产生相应变化的情况,系统提供“条目演化检索”检索功能。整个过程运用不同的线条颜色代表了保留、继承、修改、删除等多种文件间的关系。

4 结论

从两个应用项目可以看出,不论是静态知识还是动态知识,本体应用环境都能够较好地建立整个应用系统。本体应用环境的主要优势

在于:① 本体应用环境是一个完整的体系结构,不仅包括理论基础,还包括可实际应用的 API 组件。② 本体应用环境中的本体分子理论和组件可以很好地解决动态知识问题,它是对本体理论的补充和扩展。③ 本体应用环境面向实用,不要求框架中所有组件的全部实现,用户可以根据领域需求自主选择 and 组装部分组件。这种松散耦合的灵活架构能够满足大部分领域应用。

参考文献:

- [1] Berners-Lee T, Hendler J, Lassila O. The semantic web[J]. Scientific American, 2001, 284(5), 34-43.
- [2] W3C. Semantic Web Activity Statement. W3C publication[OL]. [2007-07-05]. <http://www.w3.org/2001/sw/Activity>.
- [3] OWL web ontology language reference[OL]. [2007-07-05]. <http://www.w3.org/TR/owl-ref/>.
- [4] Lassila O, Swick R. Resource Description Framework (RDF) model and syntax specification[OL]. [2007-07-05]. <http://www.w3.org/TR/rdf-syntax-grammar/>.
- [5] McGuinness D L, Van Harmelen F. OWL web ontology language overview [OL]. [2007-07-05]. <http://www.w3.org/TR/owl-features/>.
- [6] Davies J, Van Harmelen F, Fensel D. Towards the semantic web: Ontology-driven knowledge management[M]. John Wiley & Sons, 2002.
- [7] Brickley D, Miller L. FOAF: the 'friend of a friend' vocabulary [OL]. [2008-01-02]. <http://xmlns.com/foaf/0.1/>.
- [8] Knublauch H, Fergerson R W, Noy N, et al. The protégé OWL plugin: An open development environment for Semantic Web Applications [G]// ISWC. In Third International Semantic Web Conference, 2004.
- [9] Buitelaar P, Olejnik D, Sintek M. OntoLT: A protege plug-in for ontology extraction from text [G]// ISWC. Proceedings of the International Semantic Web Conference, 2003.
- [10] Celjaska D, Vargas-Vera M. Ontosophie: A semi-automatic system for ontology population from text [G]// ICON. International Conference on Natural Language Processing, 2004.
- [11] Soderland S. Learning to extract text-based information from the World Wide Web[G]// Third international conference on Knowledge Discovery and Data Mining, 1997.
- [12] Zeldman J. Designing with web standards[M]. New Riders Publishing, Thousand Oaks, CA, 2003.
- [13] TouchGraph LLC. TouchGraph[OL]. [2008-01-02]. <http://www.touchgraph.com/>.
- [14] Heer J. Prefuse: toolkit structure introduction [OL]. [2007-02-25]. <http://prefuse.org/doc/manual/introduction/structure/>.
- [15] Libick M. Who runs what in the global information grid: ways to share local and global responsibility [M]. Santa Monica, CA: RAND, 2000.

董 慧 武汉大学信息管理学院教授、博士生导师。通讯地址:武汉大学信息资源研究中心。邮编 430072。

王 菲 姜 赢 曾 杰 武汉大学信息管理学院博士研究生。通讯地址同上。

高 巾 翁丹丹 武汉大学信息管理学院硕士研究生。通讯地址同上。

(收稿日期:2009-03-04)