●董 慧 杜文华

基于本体和多代理的数字图书馆信息检索模型*

摘 要 在分析图书馆传统的信息检索机制的局限性的基础上,提出了基于本体和多代理的数字图书馆信息检索模型,并介绍了该模型各部分的作用和功能。图 3。参考文献 3。

关键词 数字图书馆 信息检索模型 本体 多代理
分类号 G250.76

ABSTRACT After analyzing the limitation of traditional library information retrieval mechanisms, the authors propose a model of digital library information retrieval based on ontology and multiple agents, and introduce the functions of components parts of the model. 3 figs. 3 refs.

KEY WORDS Digital library. Information retrieval model. Ontology. Multiple agents.
CLASS NUMBER G250.76

1 图书馆传统的信息检索机制的局限性

一般来说,信息检索机制有两方面的含义:一是检索技术,二是检索效率的评价。

传统检索技术实现的方法多采用词切分、单汉字以及词切分和单汉字相结合;检索主要借助目录、索引和关键词等方法来实现。此技术的优点是简单快捷;但缺点是无法挖掘信息之间的内在联系,检索结果不能准确、全面反映用户的需求。

传统检索效率的理想要求是快、准、全。在保证查全率(Recall)与查准率(Precision)前提下的快速 3 项指标作为对检索效果进行量化的评价标准,但是在海量的互联网上的数字图书馆信息检索上用查全率与查准率来衡量检索效率是否合适?在某些场合,高的查全率带来的成千上万条命中记录对用户实在是一个沉重的负担。

总之,传统的信息检索机制在数字图书馆中存在 3 个深层次的问题。这 3 个问题都与词汇紧密相关。这 3 个深层次的问题是:第一,"忠实表达"问题。很多情况下,用户很难简单地用关键词或关键词串

担,使用方便,易用性好。(2)通过学习、反馈等多种形式进行全过程的人机交互,有利于准确了解用户的真实检索需求,同时,在检索过程中,通过与用户进行交互,修正和调整检索策略。(3)突破了传统检索的关键词局限。仅靠关键词的出现与否进行文献取舍,易造成检出文献过多或检出文献过少的局面。(4)尤其适用目标文档中无确切查询用词,但有该词的同义词、近义词和关联词的情况。

当然,该模型也存在一定的缺陷,如后控概念词典需要人工辅助进行定期更新,新词与新的概念也需要借助人工辅助才能准确地加入概念网络等。以人工辅助方式进行知识库的更新,虽然能保证较高的质量,但耗时大,效率低,在一定程度上影响检索效率。同时,如果目标文档中有确切查询词时,查询效率不如基于关键词的检索方式。因而,下一步的改进是在检索系统设计中把基于自然语言的概念检索与基于关键词的匹配检索有机结合起来,提高检索性能。

参考文献

1 张玉峰,晏创业.基于机器学习的知识检索模型研究.图书情报知识,2002(4)

2,3 张琪玉.情报语言学基础.武汉:武汉大学出版社,2001

4 李蕾等.基于语义网络的概念检索研究.情报学报,2000(5)

5 陈定权.web 结构挖掘.情报理论与实践,2003(1)

金 燕 武汉大学信息资源中心博士生。通讯地址:武汉。邮编 430072。

张玉峰 武汉大学信息资源中心教授,博士生导师。通讯地址同上。

(来稿时间:2003-07-18)

* 本文属国家社会科学基金项目"数字图书馆相关关键技术研究"(批准号:00BTQ004)的成果。

来忠实地表达他所真正需要检索的内容,表达困难导致检索困难。第二,"表达差异"问题。人类的自然语言中,随着时间、地域或领域的改变,同一概念可以用不同的语言表现形式来表达。因此,对同一概念的检索,不同的用户可能使用不同的关键词来查询。第三,"词汇孤岛"问题。人的大脑中,概念并不是孤立存在的,它总是与其他概念之间存在各种各样的联系。在信息检索中,用户在检索一个词(表达一个概念)时,除了希望得到含该概念的文档,总是还想得到与此概念相关的其他信息,虽然这种愿望在很多情形下并没有显式地被用户表达出来。使用传统的信息检索技术,用户的这种愿望是实现不了的,因为检索返回的结果都是含用户检索词的文档,而不会涉及其他相关信息。在这种检索模式下,用户的检索词得不到概念扩展,被系统作为一个孤立的词来处理,形成了我们称之为"词汇孤岛"的问题。在这种检索模式下,如果用户要查询相关信息,他就必须再次输入相关词汇,造成这种困难的实质在于传统的信息检索技术缺乏知识处理能力和理解能力。解决问题的根本和关键,在于把信息检索从传统的基于关键词层面提高到基于知识层面。正是基于这一认识,本文提出了基于本体和多代理的数字图书馆信息检索模型。

2 基于本体和多代理的数字图书馆信息检索模型

基于本体和多代理的数字图书馆信息检索模型如图1所示。

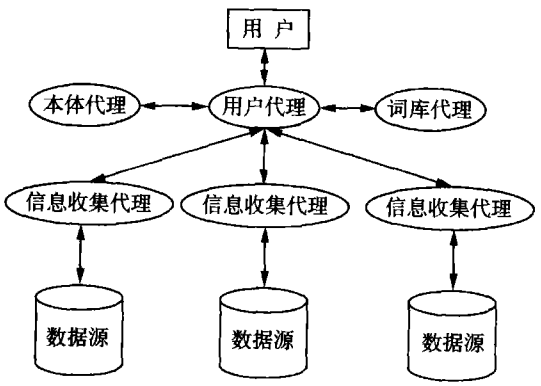

图1 基于本体和多代理的数字图书馆信息检索模型

2.1 用户代理

用户代理集界面代理和管理代理于一身,管理某个特定用户的检索情况。随着用户建模技术的运用和其推理能力的增加,用户代理能够指导用户获得他所需要的信息。用户代理也是一个管理代理。它将用户的查询任务分成几个子任务,然后将子任务分派给其他的代理。用户代理的另一个重要任务是为用户建模。在用户通过输入界面进行检索的过程中,用户代理建立用户日志文件,记下用户访问的每个步骤。如果用户代理有了用户模型,它就能够事先为用户提供更多的检索帮助。

2.2 词库代理

词库代理具有用户输入关键词之间关系的知识。它帮助用户代理辨析用户输入的关键词的语义。有关词之间关系的知识就在词林里面。本文提出的模型中有两个词库:一个是通用领域词库,一个是特定领域词库。前者如汉语同义词典,后者则只包含用在某个具体领域内的词。

概念泛化是扩大检索范围,概念具化是缩小检索范围,概念改变检索范围。在实际应用中,如果检索出来的结果太少,用户代理应该根据词库中的知识通过泛化检索范围来扩大检索结果。反之,如果检索出来的结果太多,用户代理应该缩小检索范围,而词库代理将执行具化操作。每一个操作的具体设计视所用结构而定。

2.3 本体代理

本体代理具有某个应用领域的领域知识。本体是对于"概念化"的明确表达,它研究特定领域知识的对象分类、对象属性和对象间的关系,它为领域知识的描述提供术语。本体代理运用概念之间的内在联系来指导用户代理,缩小检索范围,专注于用户要查找的主题进行检索。

2.4 信息搜集代理

信息搜集代理用于从特定的信息源来抽取信息。

3 基于领域本体的信息访问

3.1 什么是本体

本体是一个哲学上的概念,用于描述事物的本质,在知识工程及其相关的应用领域受到广泛关注。许多学科和研究领域都在使用"本体论"这个术语,但存在不同的定义。

比较通用的说法是:本体是某领域内概念的显式说明,即把现实世界中的某个领域抽象成一组概念(如实体、属性、进程等)及概念间的关系,构造出这个领域的本体。

本体论成为知识获取和表示、规划、进程管理、数据库框架集成、自然语言处理和企业模拟等研究

领域的核心。已有一些本体建立,如 Cue、Ontolingua、Microcosms、CNR、IRST、Word Net 自然语言方面的本体,TOVE、Edinburgh Plan、ARPI 等是用于规划管理方面的本体。

我们可将本体运用到信息表示、信息整合和系统开发中。本体的分析澄清了领域知识的结构,统一了术语和概念,使知识共享成为可能,使人与系统之间能够进行一致语义的交流。

3.2 本体构造的方法

建立本体的方法如图 2。

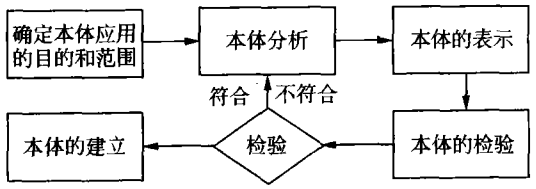

图 2 本体构造图

(1)确定本体应用的目的和范围:这是建立本体的第一步,也就是所研究的领域或任务,建立相应的领域本体论或过程本体,领域越大,所建本体就越大,因此需限制研究的范围。

(2)本体分析:定义本体所有术语的意义及其之间的关系,该步骤领域的专家参与,对该领域越了解,所建本体就越完善。

(3)本体表示:一般用语义模型表示本体。

(4)本体检验:建立本体的基本标准是清晰性、一致性、完整性、可扩展性。清晰性就是本体中的术语应被无歧义地定义。一致性,也就是术语之间关系逻辑上应一致。本体中的概念及其关系也应是完整的,应包含该领域内所有概念,但往往很难达到,需不断完善。本体应该能够扩展,在该领域不断发展时能加入新的概念。

(5)本体的建立:对所建立本体按以上标准进行检验,符合要求的可以文件形式存放,否则转入本体分析。

3.3 基于本体和多代理的信息检索模型中的知识层次

本模型中代理的知识层次分为 3 层,并且都用领域本体来表示。第 1 层是用户查询。用户查询是用户关心的一个实例。第 2 层是问题求解知识。代理用问题求解知识来决定用户查询的语义,并用与之相对应的启发策略来规范用户的查询。问题求解知识在领域专家的指导下建立。第 3 层是对特定数据的抽取知识。系统代理的知识层次和领域本体之间的关系如图 3 所示。本模型的关键知识是领域本体。领域本体定义了某领域内的对象和对象之间的关系。

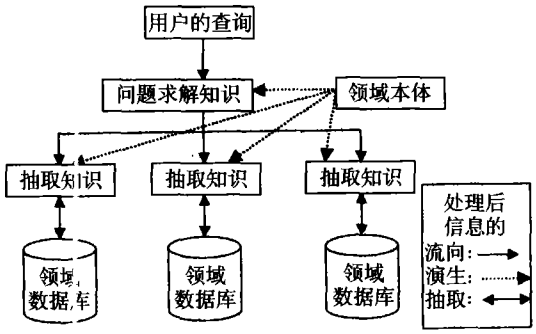

图 3 系统代理的知识层次和领域本体之间的关系

4 模型的可行性分析

国外在基于本体的信息检索方面,目前已经取得了不少进展,这些成果对于本文提出模型的可行性提供了基础。

OntoSeek 系统是一个基于内容的检索在线黄页和商品目录的信息检索系统[1]。它将本体驱动、内容匹配和丰富的表现形式有机地结合起来,在语言本体 SENSUS(从 WORDNET 演变而来)的支持下,OntoSeek 能够同时提高内容查询的准确率和查全率。

Embley 提出了一种从基于应用本体的半结构化文档抽取信息的思路[2]。他用面向对象系统模型来表示本体,用规则来识别文本并将其转换为本体中相应的角色。我们可将其思路借用到根据领域本体来从用户输入的自然语言查询中抽取信息。

Liao 提出了基于本体的检索算法[3]。我们可根据中文的特点,将其改写,使之满足中文信息检索的要求。

参考文献

1 N. Guarino, C. Masolo, and G. Vetere. OntoSeek: Content-Based Access to the Web. IEEE Intelligent Systems, 1999, pp. 70-80

2 D. W. Embley, Y. Jiang, and Y. -K, Ng. Ontology-Based Extraction and Structuring of Information from Data-rich Unstructured Documents. Proceedings of ACM Conference on Information and Knowledge Management, 1998, pp, 52-59

3 Liao M. Abecker A. Bemrdi A. et al. Ontologies for Knowledge Retrieval in Organization Memories. In Workshop on "Learning software Organization" in SEKE'99, Kaiserslautern, Germany, 1999

董 慧 武汉大学信息管理学院教授、博士生导师,软件工程国家重点实验室客座研究员。通讯地址:武汉。邮编 430072。

杜文华 武汉大学信息管理学院情报学博士生。通讯地址同上。

(来稿时间:2003-05-12)