

●张玉峰 李 敏 晏创业

论知识检索与信息检索^{*}

摘 要 知识检索与信息检索在检索概念、模型和检索语言方面既有区别又相互关联、互为促进和发展。创新知识检索方法,集成运用各种检索技术,是全面提高检索效率的根本途径。参考文献 4。

关键词 知识检索 信息检索 检索概念 检索模型 检索语言

分类号 G354

ABSTRACT Knowledge retrieval and information retrieval are different and correlated in concepts, models and retrieval languages. The authors think that innovation knowledge retrieval method and integrated applications of various retrieval techniques are major approaches to improve retrieval efficiency. 4 refs.

KEY WORDS Knowledge retrieval. Information retrieval. Retrieval concept. Retrieval model. Retrieval language.

CLASS NUMBER G354

因特网信息资源的骤增及其异构性、分布性和动态性,不断给信息检索带来新的挑战。信息检索已成为现代社会信息化和各种应用的关键。各类搜索引擎不断改进,但检索效率总是不尽人意。人们在思考和探索,如何更高层次的模拟、应用人脑的智能原理,从本质上变革信息资源检索理论、方法与技术。

实践证明,将人工智能技术与信息技术结合,发挥人工智能的作用,是一条成功的经验。例如,信息检索与知识检索的结合、数据库技术与知识库技术的结合、数据处理与知识处理的结合等。尤其是当前海量的数字信息资源为人工智能提供了一个类型齐全、内容丰富的信息知识环境,各类信息、知识与技术学习、集成和创新成为可能,而避开了与离散的物理世界打交道的许多困难。因此,科学地认识和创新知识检索方法及其与信息检索技术的结合,是现代化信息知识检索理论研究的热点。下面就知识检索与信息检索在检索概念、检索模型和检索语言方面的关联与发展,作初步的探讨。

1 检索概念

信息检索与知识检索的概念,犹如信息与知识的概念,二者之间有着本质的区别,但又互相关联、

互相交错,在使用过程中互相结合。

1.1 信息检索

所谓信息检索,就是基于信息组织形式,如字符串、结构化数据库,应用信息处理方法,如排序、数据查找、字符匹配,实现效率不高的检索。

目前,信息检索基于两种信息组织方式:线性组织和连接组织。大型数据库的基本组织方式是线性的,记录之间是线性关系,仅通过标引词索引提供相关关系,不能充分表达语义内容和语义关系,尤其是事物之间的本质关系。它通过倒排索引提供语词检索方式。网络分布式信息资源通过简单的超链连接,提供盲目的沿链式搜索方法。

信息检索综合应用布尔检索方法和基于超链的搜索技术,改进了基本检索功能,但失败于查找所有数据类型,尤其是知识内容,对精确的提问不能给出精确的回答。因此,将信息检索技术与人工智能等先进技术相结合,探索智能化、知识化的检索方法,势在必行。

1.2 知识检索

所谓知识检索,就是综合应用信息管理科学、人工智能、认知科学及语言学等多学科的先进理论与技术,基于知识和知识组织,融合知识处理和多媒体信息处理等多种方法与技术,充分表达和优化用户

^{*}本文系教育部人文社会科学研究重大项目(信息可视化与知识检索)之成果。

需求,能高效存取所有媒体类型的知识源(文本、图像、视频、声音等),并能准确精选用户需要的结果。

知识检索所涉及的知识类型很广,其中三类知识特别重要。

(1)领域知识。它是描述应用领域中客观事物的重要知识,例如学科分类知识、元数据和专业概念知识。学科分类知识可以表达事物的本质属性和事物之间的本质关联。元数据是各分布式信息源的核心内容的抽象描述。专业概念知识,例如主题词及其相互关联知识,是最基本的知识元素,用来表达与修改文献内容和用户提问,代替单纯的统计方法,可以减少统计评价关键词所产生的副作用。

(2)用户知识。它包含:用户的需求、偏好、背景知识,用户的交互、检索行为知识,以及用户对检索机制和检索结果的反馈知识。利用用户知识改进和创新检索方法,是实现面向用户的个性化检索、主动性检索的根本措施。

(3)专家的专门知识。它是关于知识组织和知识检索的显性知识与隐性知识,尤其重要的是隐性的经验知识,即启发式知识。经验知识是专家将领域知识与具体实践相结合的产物。这类知识的获取与共享,将对检索领域发挥不可估量的作用。

知识组织是高级的逻辑组织方式,充分表达知识元素的内容及其相互之间的各种关系,包括静态的语义关系、逻辑关系和动态的操作与控制关系。主要的知识组织方法有:规则、分类树、语义网络、超文本以及其它的逻辑方法和可视化方法。知识组织与信息组织有着根本的区别,但它最终要通过信息组织方式来实现。

知识检索的基本思想是,模拟扩展人类关于知识处理与利用的智能行为和认识思维方法。例如:抽象思维方法,包括逻辑推理、机器学习、概念的概括与划分、对象的分类与继承等;形象思维方法,如类比、联想等典型方法。

知识检索具有明显的优势:①实现信息服务向知识服务的转化,向用户提供潜在内容知识,以及分析、预测后的超前性领域成果或知识;②提供主动服务方式,如:主动给用户以智能辅助,主动学习用户知识并自动优化用户需求,以及主动提供个性化检索;③面向用户,检索机制的模式和界面服务方式均体现面向用户的思想,依据用户的需求及其变化,能灵活选择理想的检索策略和技术,而不是让用户机械地适应固定不变的系统逻辑,并且将繁重的知识信息存取工作从用户移向了计算机;④集成和综合

应用各类知识和各种高效的智能与非智能技术,全面提高检索效率。

2 检索模型

2.1 信息检索模型

信息检索从脱机检索到联机检索,再到网络检索,逐步发展形成了一些成熟的检索模型,包括布尔检索模型、向量检索模型、模糊检索模型和概率检索模型等。它们依据不同的检索理论,对检索元素有不同的描述。由于概率检索模型中难以获得先验数据且计算复杂,不适于大规模的网络资源,这里不加叙述。

(1)布尔检索模型

布尔检索模型是最简单易理解的模型,被大多数信息检索系统使用。它基于布尔逻辑和布尔集合论,利用关键词或其他标引词描述文献,提问是检索词的布尔组配表达式,检索方法采用精确的关键词匹配方法。

该模型易实现,计算耗费低,结构化提问语言的表达力强,但对用户构造和修改提问的能力要求较高,检索结果随用户技能而有很大差异,面对大规模网络资源出现许多问题。

(2)向量检索模型

向量模型将 N 个关键词看作一个 N 维空间。文献和提问信息用空间中的向量表示。一个文献向量由一组关键词及其对应权值表示,自动标引利用统计技术,将词在文献中的出现频率作为权值。提问向量中也可以给词加权。向量检索模型基于统计理论,用相似统计方法计算文献向量与提问向量之间的相关性。

该模型克服了布尔检索模型的缺点,许多实验证明其结果优于布尔检索模型,其相似计算结果可用于检索结果的排序输出。但它丢失了布尔提问表达式中的固有结构,不能表达检索词之间的逻辑关系。

(3)模糊检索模型

模糊检索模型基于模糊逻辑和模糊集合论。文献描述中,将概念的隶属度作为权值。提问表示为模糊表达式,执行模糊运算。它提供了对不确定信息知识的描述方法和处理方法,较接近于客观世界,但有关理论与技术尚不完善。

2.2 知识检索模型

知识检索模型,集成各类知识对象和信息对象,例如领域知识、用户知识、专家经验知识以及多媒体

信息元素等;融合各种智能与非智能理论、方法与技术,例如推理、机器学习、数据挖掘、智能搜索、智能演算以及多媒体信息处理等;实现知识检索,例如基于语义内容的检索、基于知识结构的检索、基于推理与学习技术的检索、基于专家与用户知识的智能搜索,以及分布式多维检索等。这里侧重描述模型的知识对象组织和知识处理方法。

下面介绍三种知识检索模型:分类检索模型、认知检索模型和分布式检索模型。它们采用不同的知识组织模式,应用不同的人工智能技术和知识处理方法,尚不成熟。

(1) 分类检索模型

分类检索模型的核心思想是数据抽象,利用事物之间最本质的关系来组织资源对象。概念逻辑与知识分类思想是知识标引和知识检索的理论基础。

知识组织结构是类层次结构。知识对象由静态和动态的知识元素构成,对象之间存在类属关系和其它语义相关关系。

用户提问对象同知识库中知识对象的结构相同,可包含用户的多维异构需求信息。

该模型综合应用类结构的查找方法和对象归类方法,实现对象或概念的检索。对象之间的等级关系,可提供快速的自顶向下的查找策略。利用继承原理,可实现自底向上的检索。其他语义相关关系,如参照关系,能实现横向的搜索。

这种模型的最大优点是语义继承特性,便于知识资源的共享,提供多途径、多方位的多种检索方法,充分表达用户的多维组合需求信息。该模型较适于领域知识和用户知识等大规模知识源。它需要综合利用自动化分类方法和手工分类方法,关于大规模资源的存储和动态维护已由计算机实现。

(2) 多维认知检索模型

认知检索模型的理论基础是人工神经网络(ANN - Artificial Neural Network)。它以模拟人脑的结构和神经系统为目标,将资源组织为语义网络结构,由结点和链接组成。结点可以表示概念或知识对象,链接表示对象之间的各种语义关系和动态操作关系。

该模型利用启发式知识和传递激活技术,形成一种认知式的语义推理模式,实现知识库的联想搜索。

该模型利用学习规则和相关反馈技术来修改、完善提问和知识库,不断改进检索结果。它适合于专业概念知识库的检索,用于扩充和优化用户需求,

还可以帮助用户区别词的多义。用户可以充分描述概念和概念之间的关系,查找精确的文献内容和执行语义链的导航搜索。它不适于大规模的知识源。

(3) 分布式检索模型

对于网络异构分布式信息资源,分布式检索很重要。该模型要求建立元知识仓库,集成和存储各类元知识。它综合应用分布式人工智能、神经网络、智能演算、并行推理、机器学习等技术,评估各类资源与用户需求的相关性,选择最好的知识源和数据库集合,分别执行并行检索。最后,它利用聚类、综合分析与学习等智能处理方法,产生全局一致的、有效的检索结果。

元知识仓库(Metaknowledge repository)包含描述各类数据库内容的元数据、用户与专家的知识、各类控制知识、各种优化的分布式检索算法和协作调控协议等。

3 检索语言

3.1 关键词

关键词是表示文献主题内容的基本单位,是最简单的自然语言。信息检索模型都将关键词作为描述文献和提问的基本元素。

关键词能自动抽取,是实现主题检索的重要途径,简单易用,对于信息检索和知识检索都是不可缺少的。

问题是只有简单的关键词,还不能充分表达各种丰富的概念,尤其是概念之间各种语义关系,因此检索效率很低,不能满足用户的要求。

3.2 自然语言

自然语言处理一直是人工智能的一个核心研究领域。自然语言处理技术是知识获取、知识表示技术的基础,也是充分发挥用户的主动性、提高检索效率的有效途径之一。

自然语言处理已成功的应用于全文检索系统,用户可以输入自然语言的字符串,系统执行模式匹配,获得较精确的检索结果,但速度慢,不适于大规模的信息集合。

自然语言处理更多的用于自然语言接口,使得人一机交互能模拟人—人交互的认知功能,实现用户与系统之间的更充分的互相交流、互相理解、互相学习、互相协作,是理想的交互语言。

自然语言处理的自动化难度很大,自然语言提问具有一定的模糊性,系统需执行一些语义检查。

3.3 知识表示语言

超文本标识语言 HTML 使得互联网高速发展。它面向数据显示,提供固定的信息描述格式和页面的简单链接方式,检索效率极低。一种新型的面向数据的语言 XML (eXtensible Markup Language),支持字符串和结构化的数据,提供多层次的链接功能,允许用户动态修改信息,明显改进了信息检索效率。随之,许多面向知识的元数据语言应运而生,它们都建立在 XML 之上,如资源描述框架 (RDF—Resource Description Framework)。因为它们都是通用的、不精确的、面向整体文献 (而非文献的知识元素) 的描述方法,且没有协议来表示基于逻辑的特征,限制了它们的描述能力,不能满足精确的、灵活的基于知识的检索。

知识表示语言是一种高级的逻辑表示语言,可以表达文献的知识内容及其相互之间的语义关系和逻辑关系,支持逻辑推理和基于知识的检索。常用的知识表示语言有:规则、谓词逻辑、框架和语义网络等。例如,概念图是一种语义网络结构,可以让用户表示详细的文献知识,只要设计一些简单的符号加以限制即可。

知识表示语言可以结合 HTML 和 XML 等语言一起使用。例如,可以设计一种符号,对 HTML Web 文献作注释时,在 HTML 超链标记中嵌入有关知识描述,如属性值对、可视化的图形或自然语言句子等。用这种方法,系统可以集成知识库和其他关联的文献集,并可执行信息检索和知识检索。

3.4 理想检索语言

理想检索语言,应集合各种语言之长而避其短,应主题语言与分类语言兼容、自然语言与人工语言并用,具有以下特征:

(1) 表达能力强,语义精确,简明易用,通用;可以表示任何可存取信息,如结构化、半结构化、非结构化信息,多类型、多媒体信息;还能充分表达各类静态与动态知识;

(2) 提供自然语言、人工语言 (包括知识表示语言) 及可视化的交互方式;

(3) 支持知识检索和信息检索功能及多种检索技术;

(4) 具有主动的自学习、自修改能力。

检索领域的当务之急是,综合利用信息管理、人工智能等多学科理论,探索知识检索的理论与方法,集成运用各类检索模型、方法与技术,实现不同级别、不同颗粒的信息检索与知识检索。这是全面提高检索效率的根本途径。

参考文献

- 1 张晓刚,李明树. 智能搜索引擎技术的研究与发展. 计算机工程与应用, 2001 (24)
- 2 张玉峰. 智能情报系统. 武汉: 武汉大学出版社, 1991
- 3 Philippe Martin and Peter W. Eklund. Knowledge Retrieval and the World Wide Web. IEEE Intelligent Systems, 2000
- 4 F. Wiesman, Arie Hasman, H. j. van den Herik. Information retrieval: an overview of system characteristics. International Journal of Medical Informatics. 1997

张玉峰 武汉大学信息资源研究中心教授、博士生导师。通讯地址: 武汉市。邮编 430072。

李 敏 武汉大学硕士研究生。通讯地址同上。

晏创业 北京大学博士研究生。通讯地址: 北京。邮编 100871。

(来稿时间: 2002-08-19)

欢迎订阅《中国图书馆学报》

《中国图书馆学报》是文化部主管、中国图书馆学会和国家图书馆主办的国家级图书情报学专业期刊,被评为中国优秀图书馆学期刊、全国中文核心期刊、中国期刊方阵期刊和国家期刊奖百种重点期刊。

《中国图书馆学报》以广大图书情报工作者和图书馆学情报学专业师生为主要对象,以开展图书情报学学术研究和交流为宗旨,是发表学术研究成果、交流学术思想的专业学术刊物。其任务是在“百花齐放、百家争鸣”方针指引下,开展学术讨论,提高学术水平,促进我国图书情报事业发展。

2004 年征订工作即将开始。全国各地邮局均可订阅,国内代号 2-408。中国国际图书贸易总公司负责国外发行,国外代号 Q184。2004 年每期定价 10 元,全年 60 元。