我国智能化信息检索发展及研究现状

韩娇红
(安阳师范学院计算机与信息工程学院,河南 安阳 455002)

[摘 要]从我国对智能化信息检索的发展、智能信息化检索的基本理论、智能化信息检索实现的可能性和技术要求、智能化信息检索面临的问题和解决的思路、智能化信息检索技术发展趋势 6 个方面的认识入手,探讨了我国智能化信息检索研究现状及趋势问题。

[关键词]信息检索 智能化 检索技术 发展趋势
[分类号]G254.91

1 引言

社会进入信息时代,不仅信息的数量急剧增长,而且信息检索的对象和信息需求的主体也发生了极大变化。信息检索的对象从相对封闭、集中管理的信息内容扩展到开放、动态、更新快、分布广泛、管理松散的 Web 内容;信息用户由原来的情报专业人员扩展到包括商务人员、管理人员、教师学生、各专业人士等在内的普通大众,他们对信息检索从结果到方式提出了更高、更多样化的要求。目前,信息检索已经发展到网络化和智能化阶段,适应网络化、智能化以及个性化的需要是目前信息检索技术发展的新趋势。其中智能化信息检索成为研究的热点。

2 智能信息检索发展概述

2.1 智能信息检索研究的几个阶段

2.1.1 发展初期及形式(1985~1987)

网络化的逐步形成日益改变了人们的工作和生活方式,尤其被称为人类三大尖端技术之一的人工智能技术的出现和发展,给人类社会带来了强烈的活力,传统的信息检索在理论和实践方面因此受到了巨大的冲击,表现出许多自身难以克服的弊端,国外智能信息检索就是在这种背景下提出的。在我国,信息工作者也不甘落后,从 20 世纪 80 年代中期开始了对智能信息检索的研究。

2.1.2 发展高峰期(1988~1991)

人工智能经过多年的发展已日趋完善。国外在 20 世纪 90 年代初已达到了高峰,这也为我国智能信息检索提供了基础。从国内来看,这也是我国情报学发展的黄金时期,大量情报学文献在此期间问世,其中还包括大量智能情报检索研究的文献。

2.1.3 逐步回落时期(1992-)

人工智能信息检索是一项难度很大的课题,首先要解决电脑思维与人脑思维有机结合的问题,在人工智能没有突破性进展的情况下,其应用领域也不会有大的进步。计算机表现出的局限性使信息检索智能化发展不可避免地受到影响。我国对智能信息检索的研究开始原地踏步,甚至回落,从高潮走进了低谷。

2.2 我国智能信息检索研究现状

目前,我国的智能信息检索基本处在理论认识与实验阶段。通过传统信息检索系统与智能信息检索系统的比较分析,人们对智能信息检索系统已经有了理性的认识,对智能信息检索系统的基本结构有了较清晰的认识,对智能信息检索系统实现方法和关键技术逐步了解,看到了智能信息检索系统实现的可能性,同时也面临着许多困难和问题。

3 智能信息检索的基本理论

3.1 智能信息检索的概念

信息检索(Information Retrieval),通常指文本信息检索,包括信息的存储、组织、表现、查询、存取等各个方面,其核心为文本信息的索引和检索。智能检索则是把现代人工智能的技术与方法引入到信息检索系统,使后者具有一定程度的智能特征,在更高的层次上实现其功能。智能化信息检索的目的是使信息检索系统"理解"文件包含的信息内容和用户的信息需要。它在对内容的分析理解、内容表达、知识学习、推理机制、决策等基础上实现检索的智能化。

3.2 智能信息检索系统应具备的能力

智能信息检索系统应具备以下 3 种能力:①智能信息检索系统是建立在大规模的知识库基础之上的,它有一个强大的推理系统支持,能用自然语言而不是规范的主题词与检索者交流的计算机系统。此系统能在已知信息的基础上,推理分析出系统没有明显表示出来的信息。此外,系统自身还具有学习和自适应能力。②智能信息检索系统在具备知识库和推理机制的同时,强调智能信息检索结果应是用户能够直接加以利用的信息,与传统信息检索为用户提供的文献线索相区别。前者可以免去用户再去查找文献的重复活动。③智能信息检索系统的智能因素不应仅仅定义在检索的执行过程中,还应体现在提问模型的形成过程中,即根据用户对问题的描述,借助与知识库相关的知识,推断其真正需求,形成提问模型。

3.3 智能信息检索的系统结构

一般来说,智能信息检索系统由知识库、文本处理和智能接口 3 部分组成。①知识库部分 知识库是智能检索的核

心。它又由知识库系统、数据库系统和检索推理系统 3 个子系统构成。②文本处理部分:文本处理系统就是利用计算机自动处理自然语言形式的文本输入。它利用知识库中的语言学知识、科学知识和其他知识,对文本进行语法、语义分析界定,从内容上理解文献所论述的主题,并把它们表示成知识库中的知识单元和数据库中的数据元素,不断地丰富知识库和数据库。③智能接口部分:智能接口是用户与系统之间的通道。它的主要功能是对自然语言进行查询和处理,并作为智能终端建立用户兴趣档案,加工提取结果。

4 实现的可能和技术要求

4.1 实现的可能

人工智能技术中的机器感知(知识获取)、机器思维(知识处理)、机器行为(知识利用),其核心是知识。实现知识的形式化描述,从知识的获取、表示、存储、组织、管理、推理直到智能化研究一直是人工智能研究的主要方向。

信息检索现在虽然还没有达到知识层次上的加工处理,但它至少已通过间接的途径实现了对知识的处理,如智能化信息检索系统把信息源作为知识的集合,而把信息源通过适当的方式加以标引,其目的也在于通过这些标引词来表达信息源中的知识点,并为用户的信息需求提供相应的知识辅助。智能化信息检索系统的目标就是真正达到在知识语义层次上进行信息服务。由此看出,知识是 IR 与 IT 共同的研究对象,而对知识的获取、加工、处理、提供利用则是两者共同的目标。目前的智能技术主要包括人工智能技术和人工神经网络技术,其中人工智能技术(AI,Artificial Intelligence)主要研究如何利用计算机软、硬件模仿、延伸、扩展人类智能理论方法和技术。而人工神经网络技术(ANN,Artificial Neural Network)则更注重对人脑结构的模拟。实际应用中往往可以通过结合 AI 与 ANN 共同完成智能任务。AI 长于知识的逻辑推理,它以一套完整的推理系统为核心,对知识进行组织、再生和利用;ANN 的长处则在于对复杂知识的结构化组织,通过分布式计算、并行推理以及例子学习来实现智能化处理。ANN 是模拟神经元结构,决定了它具有高度容错能力。ANN 的研究重点在于模拟和实现人的认知过程中的感知过程、经验形象思维、分布式记忆和自组织学习过程,而 AI 是符号处理系统,侧重于人的逻辑思维。这两者的结合为基于知识的智能化的信息检索提供了可能。

另外,随着计算机软、硬件设备性能的提高和智能通讯、网络技术的深入研究,人工智能在自然语言理解、知识获取、表示和推理等方面的研究进展,以及信息检索领域对智能化的努力,为两者的结合提供了强大的技术支持和广阔的应用空间。

4.2 技术要求

用户知识的自动获取技术。用户知识通常包括用户信息需求和用户背景知识等。通过在用户终端上运行一个监视用户的接口 Agent,由它来监视用户信息搜索与浏览过程,将用户在 WEB 浏览时的相关信息不断传给远端服务器,服务器再将信息进行整理、组织并从中分析出用户的信息偏好,服务器根据用户信息偏好进行新的信息推荐。

特征提取技术。在智能检索系统中读取文档,分析其结构并从中提取对用户查询有益的索引数据。

机器学习技术。包括基于解释经验的学习、基于事例的学习、基于概念的学习、基于类比的学习、基于神经网络的学习等。其具体的执行是先让一个智能 Agent 带有最小的背景知识,然后通过几种方法学习用户的行为:一是观察用户,找出规律;二是用户反馈(直接或间接);三是用户训练,直接给出例子;四是询问其他 Agent。这样即使 Agent 不熟悉某个用户的习惯,但经过一段时间的学习,它会逐渐了解用户的工作习惯,并逐步接替用户的工作。

推送技术。推送技术最基本的形式是通知,针对这种服务,用户可以控制其通知形式与时间间隔。另一个是提要技术,用户以关键词、日期、数值、比较规则以及其他查询条件查找信息。提要可以实现查看 WEB 页或其他信息源,寻找需要匹配的信息,并向用户传递信息。第三种是自动拉出,提供一种可供用户常查看的 WEB 页。自动推送需要用户终端有特殊客户机软件,定期发出更新请求。

5 面临问题和解决思路

5.1 面临问题

5.1.1 智能技术本身的不成熟

人工智能技术本身还有许多不完善的地方。主要体现在两个方面:①知识的获取与表示。其中较难解决的问题就是如何把复杂多样的专业知识系统化。此外,如果把人工智能技术应用到一个多学科综合的检索系统中,如何辨别某个多义词当前的具体含义,如何辨析用户特定的需求,这些都有待于继续研究。②受自然语言处理技术方面的局限。要想使计算机准确地分析、表达并传输知识,就必须使计算机具备理解自然语言的能力。目前对自然语言的处理,虽然已从语法阶段上升到语义阶段,但对自然语言的理解能力还限制在一些规范的语句和语法范围内,这就决定了智能信息检索系统所能具有的智能化表达程度。

5.1.2 信息检索系统本身的障碍

信息检索系统是一个复杂的系统,检索过程本身存在着以下难题:①信息检索系统所面对的用户来自不同专业领域,知识层次也各不相同,要使计算机对其进行合理定位是一个难题。②信息检索系统涉及的专业知识丰富,将诸多知识形式化较为困难。③信息检索专家系统不易建立。不仅这些专家的经验和技术很难准确地表达出来,而且不同的检索

专家很可能对同一问题持不同的观点,这对专家系统的建立提出了难题。

5.2 解决思路

5.2.1 解决知识表达问题的思路

知识的获取和表示问题是智能化信息检索的一个难题,但是知识库是智能检索的核心,如何建设知识库,关键是如何把复杂多样的专业知识表达描述出来。在我国,不同的学者从不同的角度去探讨这个问题,有人认为语料库作为处理自然语言的方法较好,可用来构建语义网或采用本体论建设知识库。目前,随着网络信息的多样化、网络数据库的异构化,本体论越来越受到了计算机界的重视。在协助智能体对因特网上的各种信息进行领域分类,在智能化的规范用户信息检索和信息整合方面,本体论的知识发挥着重要作用。由于本体能刻画事物之间的内在联系,借助于本体,可以使检索的信息更能满足用户的需求。所以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。一旦建成基于本体论的知识库,本体论将提供一个内容丰富和现代的框架以实现术语的规范、服务和管理。如果与基于网站的搜索工具相结合,将会十分有益于资源的检索,不仅可以为特定用户提供其所查询的特定文件,还可提供与兴趣主题可能有关的其他资源。这种额外的功能不仅会显著提高基于网站的搜索引擎的范围,而且还能改进用户对网页上信息资源感兴趣的方式。

5.2.2 自然语言处理的思路

语言学方法。根据可计算性理论,任何一个自动机的运算都是按一定程序、分步骤和相继作用在离散对象之上所完成的,而这些对象又是以线形序列相邻接的排列组合所构成的,而自然语言的 3 个特征——离散性、序列性、邻接性使其具备了"可计算性",为自然语言的处理奠定了物质基础。对自然语言处理的方法有语言学方法、人工神经网络法等。与建立语料库的思路相似,采用语言学方法,在相当长时间里,语言学的任务是建立一个高度集合的语法系统,来解释句子的生成与理解。当这一语言学理论与计算理论相结合时,产生了形式之法。形式之法由一套有穷的规则结合所组成,其作用是生成并接受所有符合这些规则的语句。

语料库方法。语料,又被称为素材,是自然发生的语言材料的集合。而语料(Corpus)是一个由大量在真实文本经过词法、句法、语义等多层次加工形成的语言材料库。这些加工的方式包括在语料中标注各种记号,标注的内容包括每个词的词性、语义项、短语结构、句型和句间关系等。随着标注程度的加深,语料库逐渐熟化,成为一个分布的、统计意义上的知识源。语料库本身不能直接应用于自然语言处理中的句法或语义分析,但因为语料库包含了语言或者语言变体的词汇、语法结构、语义和语用信息,为语言学的研究提供了无穷无尽的资料来源,是计算机对文本进行各种分类、统计、检索、综合、比较等研究的基础,可以帮助语言学家揭示语言的词汇、语法、语义和语用规律,由这些语言学的规律汇集成词法、语法、语义词典或知识库等文本分析工具,然后利用这些工具进一步对其他大量新文本逐词标注词性,划分句子成分,进行语义标注等。

6 智能信息检索技术发展趋势

互联网上利用搜索引擎为检索手段,使用网络信息资源自动采集机器人(robo)程序(也称网络蜘蛛、爬虫软件),动态访问各站点,收集信息,建立索引,并自动生成有关资源的简单描述,存入数据库中供检索。但这种机器人程序的查准率有待提高。于是元搜索引擎(又称多元搜索引擎或集成搜索引擎)成为网络检索的后起之秀,是多个单一搜索引擎的集合。它没有独立的数据库,主要依靠系统提供的统一界面,构成一个一对多的分布式且具有独立功能的虚拟逻辑机制。以上两者都不能提供用户直接利用的信息资源,且查准率有待提高。网络智能检索成为目前研究的热点,其包括智能搜索引擎(Intelligent Search Engine)、智能浏览器(Intelligent Browser)、智能体(Agent)等。智能搜索引擎可以预期用户的需求,并可有效地抑制关键词的多义性。比较成功的智能搜索引擎有 FSA、Eloise 和 FAQFinder。智能浏览器是基于机器学习理论设计的智能系统,经过训练后,可成为某个领域中熟练的搜索专家。两个比较成功的实验原型是 WebWatcher 和 Letizia。智能体是一个具有控制问题求解机理的计算单元,网络中的智能体通常是一个专家系统、一个模块等。它在经用户指导后,可在不用用户干预的情况下,找到所需信息。有些智能体使用神经网络与模糊逻辑而不是关键词来识别信息的模式。例如 BrowerBuddy 是一个基于规则的智能体。

当前基于 Agent 的智能信息检索是信息检索技术研究的热点。智能代理(Intelligent Agent,简称 IA)技术始于 20 世纪 80 年代,是人工智能技术的一个重要研究领域。进入上世纪 90 年代后,随着因特网的广泛使用及其相关技术的飞速发展,围绕因特网展开的智能代理技术研究取得了很大的进展,它不仅成为人工智能研究的热点之一,也是信息技术最前沿的代表。智能代理最先由美国麻省理工学院研制开发。目前,国外从事智能代理技术研究的不仅有大学、研究机构,还有 Apple、IBM、微软等诸多信息技术公司,并且有些智能代理产品或嵌入智能代理技术的产品已经投入使用。这些情况表明发展智能代理技术是一个趋势,它将是克服现有网络信息检索问题的有效手段。

目前信息检索技术正朝着多功能和智能化方向发展,随着自然语言处理、自动分词、自动标引、自动文摘、自动分类、自动翻译等技术的进一步发展,信息检索技(下转第 63 页)

享系统的建立,各校都建有自己的特色教学资源库、精品课程库。为充分发挥南通地区高校图书馆的综合优势,积极探索切实可行的共建、共享模式,推进教育文献信息资源的开放和共享,提升各成员馆服务教学科研的能力,迫切需要建设联盟机构知识库。

4.2 南通高校联合体联盟机构知识库的构建模式

在南通高校教学联合体中,南通大学是唯一一所本科综合性大学,人力和物力都具有绝对优势,技术开发力量雄厚。南通高校联合体机构联盟知识库采用的是集中模式,中央集中存储器放于南通大学,由南通大学牵头,组织其他 5 所学校共建。选择集中式模式的主要原因是,首先南通高校联合体目前还没有一个成员建有机构知识库,因此可以从顶层设计,选择统一软件,制定统一政策,有利于集中管理。其次,南通高校教学联合体都传承张謇教育思想,有相同的历史渊源,有地方特色的地方文献资源,有合作的历史。再次,地理位置相对集中,组织交流方便,有利于资源的开发和利用。

5 结语

建设机构知识库,利益与挑战并存。笔者认为,在相当长一段时期内,联盟模式能够为知识库的发展做出积极贡献。而这种模式对于资金、人员、技术、资源都相对匮乏的国内知识库来说,无疑具有很大的借鉴意义。如何借鉴国外机构联盟知识库成功的经验来指导正在兴起的国内知识库的建设,将成为我们需要进一步探索的问题[10]。

参考文献:

[1] 渠芳.高校教学联合体机构知识库联盟建设研究[J].情报理论与实践,2010(11):83-85.

[2] 万文娟,吴高.高校教学联合体机构知识库联盟建设研究[J].国家图书馆学刊,2010(4):31-35.

[3] 邓君.机构知识库建设模式研究[J].图书情报工作,2010(6):112-116.

[4] 曾苏,等.机构知识库联盟发展现状及关键问题分析[J].图书情报工作,2009(24):106-110.

[5] 王颖洁.国外机构知识库运行模式分析[J].当代图书馆,2008(4):58-61.

[6] 陈淑珍,等.我国大学机构知识库建设的模式选择与实施策略[J].图书馆杂志,2009(8):52-54.

[7] 张东华.高校机构知识库的内容构建与管理模式研究[J].情报科学,2009(6):836-838.

[8] http://gxlh.ntu.edu.cn/.

[9] http://222.192.60.12/html/jalis/20090618153902.htm.

[10] 常唯.机构知识库:数字时代一种新的学术交流与共享方式[J].图书馆杂志,2005(3):16-19.

朱志伯 1961 年生,副教授。研究方向:知识管理、数据挖掘。

吴海霞 1965 年生,副研究馆员。研究方向:知识管理。

(收稿日期:2011-09-05;责编:杨新宽。)

(上接第 51 页) 术必将日益走向成熟与完善。

7 结束语

人工智能技术的发展是时代对社会智能化需求的体现,而人工智能与信息检索的结合则是人们对信息获取智能化的有益尝试。在信息检索系统中纳入人工智能技术将使传统的信息检索系统具有更好的用户界面、更高的检索效率和更丰富的检索手段。人工智能技术的引入正在使传统的信息检索系统发生巨大的变化。以两者作为结合点的智能信息检索系统也将随着这两方面研究的不断发展而更加完善强大。

参考文献:

[1] 师东生.基于自然语言理解的智能化多媒体信息检索系统研究[J].微型机与应用,2011(6):6-10.

[2] 宋喆,初广丽.基于 Multi-Agent 的个性化信息检索模型结构体系[J].图书馆学研究,2011(2):62-66.

[3] Liu Ying,Tang Yonglin,Zeng Yuan.A study on improving information retrieval effectiveness for scientific and technicalnovelty retrieval [C].Proceedings of International Forum onTechnological Innovation and Competitive Technical Intelligence'2008,2008:338-347.

[4] JAIN P.Intelligent information retrieval[C].SETIT 2005 3rd-International Conference:Sciences of Electronic,Technologiesof Information and Telecommunications,2005(3):27-31.

[5] KANNAN R.Topic map:an ontology framework for information Retrieval[C]. Proc.of National Conference on Advances in Knowledge Management, 2010:195-198.

[6] 肖艳华,邵世煌.一种基于本体论的 Internet 信息个性化检索系统的 Agent 实现模型[J].微计算机信息,2003(6):77-78.

[7] 何儒云,汤艳丽.智能化信息检索研究[J].图书馆,2003(3):34-37.

韩娇红 女,1971 年生。硕士,馆员。研究方向:信息资源管理、知识管理。

(收稿日期:2011-09-18;责编:张欣。)