中国图书馆学报(双月刊)2001 年第 3 期
ZHONGGUO TUSHUGUANXUEBAO

●焦玉英 晏 凌

网上主动信息服务系统的模型及其实现技术

摘 要 通过比较网络信息服务的“拉”(pull)技术和“推”(push)技术,给出网络条件下的主动信息服务界定。描述了由用户中心系统、信息中心系统、信息映射系统和信息安全系统组成的网上主动信息服务系统模型,分析了系统实现的关键技术。参考文献 5。

关键词 因特网 模型与技术 主动信息服务技术
分类号 G250.72

ABSTRACT By comparing the “pull” and “push” technologies of network information services, the authors define active information services in the network condition, describe a model of active network information service system consisting of user-centered system, information-centered system, information mapping system and information security system, and analyze key technologies for the realization of the system. 5 refs.

KEY WORDS Internet. Model and technology. Active information service technology.
CLASS NUMBER G250.72

1 界定网上主动信息服务系统

在非网络条件下,传统图书情报机构的各类型服务,如代查、代译、定题服务等,均属于被动服务方式。而在网络环境中,众多的 ISP(因特网服务提供商)、ICP(因特网内容提供商)、SE(Search Enginee:搜索引擎)以及门户网站使用户可以直接从网上获取信息,这与传统服务方式相比有较高程度的主动性,但从根本上讲还是被动服务方式。因为,它使用的技术从信息的传输方式看,是由用户通过 Browser(浏览器)向 Server(服务器)发出服务请求,Server 在所拥有的信息资源中进行查询处理,把处理结果传回 Browser 所在的计算机来实现的。在“拉”技术中,主动的一方是 Browser,Server 只是根据 Browser 的请求,被动进行信息发布。

以“拉”技术为基础的被动信息服务系统,主要存在两大致命弱点:一是 Browser 在不知主机地址只知所需主题的情况下,必须在成千上万的信息中搜索,不仅效率低,响应慢,还使信息查询的结果量大而且准确率低;二是假使一条信息被 N 个用户所需求,则它在因特网上要传输 N 次,这样就大大降低了网络效率,随着用户及信息量的增加,必然会造成网络堵塞。一旦某种信息资源的 Server 成为“热点”时,它将同时受到网上许多个 Browser 的访问,即并发访问量过大,这时就可能造成 Server 主机的崩溃、瘫痪。因此,借助人工智能等计算机技术,结合图书情报工作多年的信息处理经验,使目前的网上信息更加主动,在更高程度上满足用户的特定信息需求显得十分迫切。

网上主动信息服务系统是以“推”(Push)技术为基础的。所谓“推”(Push)技术,是与“拉”技术相对的,就是 Server 根据事先规定的设置文件,而不是根据用户即时要求,主动向 Browser 递送信息的技术。它与“拉”技术最大的不同是在“推”技术作用下,用户与信息之间的关系的改变:“拉”技术条件下是用户找信息;“推”技术条件下是信息找用户,即用户不必进行任何信息检索操作,就能方便地获得所需信息,更新后的信息被随时推送给用户。网上主动信息服务系统拓展了网络信息组织与服务技术,其频道概念,使用户浏览 Web 只需在频道之间进行切换,Web 内容将组织成一个个频道对外播出。该系统使信息服务机构除了被动地面向整个网络用户服务,还能从技术上主动锁定一批特定用户群,为他们提供网上专题信息服务,这不但提高了信息服务的效能,节省了用户在网上漫无边际查询信息的时间,还减少了网上部分无效信息的流通量,节省了宝贵的带宽资源。这无论是对网络管理还是对因特网自身的健康发展都有

积极意义。

2　基于"推"技术的网上主动信息服务系统的模型

本文对网上主动信息服务系统的设计,考虑到该系统是"被动中的主动"这一特点,旨在实现用户"拉"基础之上的信息"主动推送"。这样,使系统的设计更具有实用性和可行性。因为只有用户先在网上用"拉"的方式,让信息提供方被动地了解用户的信息需求,才能更好地分析用户,更主动向用户提供真正有价值的信息。该系统是一个智能化的高效信息系统,能使信息在正确的时间到达正确的地方,让用户能迅速感知信息的变化并做出响应,表现在:①智能化。该系统中的信息是高速流动的,不是停留在一个地方等人去寻找,而是带有目标的,采用新的分类形式,主动寻找合适的信息用户。②高效性。由于系统中的信息能主动寻找用户,信息的有效传递率会大大加大,信息利用率极大提高;用户直接面对的信息量可以得到控制,信息的商业价值得以充分发挥,避免了垃圾信息对网络资源的大量占用。

网上主动信息服务系统由4个子系统构成。

2.1　用户中心系统

用户将要在这里完成注册,表述自己的信息需求,被统计分析、被管理,只有这样Server才能够了解相对真实的有价值的用户信息,便于做成一个有效的电子身份证,向用户提供主动及时的网上信息服务。该子系统由3个功能模块组成。

2.1.1　交互机

交互机负责与用户的交互,接受其请求并向用户Browser推送其所需信息。交互机及时地通知用户网上更新了的信息,同时负责显示本地信息库的相关内容,并定期地刷新界面以反映本地信息库的动态变化。交互机实际上兼有两个功能:与用户交互,让用户尽可能地表达自己真正的信息需求;机器学习。

交互机对查找用户真正想要的信息起着非常重要的作用。它不断学习用户对某方面信息的偏好,然后建立并维护一个 用户专用词典以便根据用户的兴趣对查询模板进行特殊的解释。用户的请求包含多义性词汇时,它将弹出一对话框与用户交互以便将具有一般意义的请求缩小至具体领域,这样,就可以生成精确的查询请求。经过足够时间的验证,交互机会自动利用用户的特殊解释来指导信息的搜集。交互机提供一种手段让用户能够直接把某些特殊含义的词加入到用户词典中。为了发现用户对信息的需求,交互机监视用户的操作,当它发现用户选择并将其感兴趣的某些信息加入本地信息库的某个类中时,即利用相似性分类技术来分析这些信息并通知信息搜索机器人在因特网上搜集类似的信息。

2.1.2　语法分析器

系统用户可以利用自然语言来表述信息需求。用语法分析器对这些自然语言进行分析并生成用作信息搜集的内部查询模板。首先,语法分析器将输入的字符串分解成几个部分,每个部分可能包含几个单词,单词间的关系是通过查阅词典来分析。语法分析器可使用多个词典。如果输入的字符串中含有词典无法解释的词,则通过交互机由用户来定义,并把这些词的定义加入到用户专用词典中,以便下次使用。输入的字符串被成功分析后,一个内部查询模板就被生成并传送给信息搜索机器人,进行信息搜集。

2.1.3　词典

系统中的词典分通用词典和专用词典两种。一个与专业领域无关的通用词典包含了代词、连词、冠词和其他与分类无关的词。任意数目的专用词典包含了各个不同领域的重要术语。词典在整个系统中有重要作用,它的质量是系统质量的关键,决定着用户对信息服务的满意度及系统的智能化程度。

2.2　信息中心系统

信息中心系统负责搜集信息,对信息进行分类整理,确定标准,把个性化信息标准设立出来,使大量信息遵循这样的标准进入信息系统。该子系统由以下4个功能模块组成,共同完成对网络信息的再组织。

2.2.1　本地信息库

本地信息库建立在关系数据库系统的基础上,它将创建一个数据模型来组织从网上搜集到的信息。存放在本地信息库中的信息是分类组织存放的。信息中心系统负责维护本地信息库的动态变化。其网上信息挖掘机可以实时检查网上与本地信息库中相关的信息是否已被更新或删除,并把变化情况通知本地信息库。当本地信息库的维护信息,如分类结构等发生变化时,信息中心系统应把这些变化通知用户中心系统的交互机,以便更新用户交互界面。

网上主动信息服务系统提供了对本地信息库的多种操作:它能接受查询请求并检索所需的信息;系统用户可以向本地信息库添加信息、删除本信息库中不再需要的信息等。

结合主动信息服务系统的实际情况,本地信息库有3种方式来获取建库信息:①利用信息搜索机器

人,自动发现及搜索因特网上的信息资源。②通过镜像软件,将各类用户经常访问的网站内容镜像到本地,并进行全文检索,提供全文检索服务,使用户不离开系统,就能得到某些网站的内容。③有的用户希望将他们所了解的信息提交给系统,与其他用户共享。

2.2.2　网上信息挖掘机

网上信息挖掘机是一个小型计算机系统,它通过对大量已知数据样本进行综合分析,得到数据对象的内在特性,并以此为依据在网络中进行有目的的信息提取,其中涉及归纳学习(Inductivelearning)、机器学习(Machinelearning)和统计分析(Statistics)等人工智能技术。在现有技术下,本系统的网上信息挖掘机仅能对万维网上的文本信息进行挖掘。

网上信息挖掘机一般由特征提取、源信息采集、特征匹配 3 个部分组成。特征提取负责根据一定的算法和策略从现有的样本文档中提取出其内在的特征,即进行挖掘目标的特征提取。源信息采集负责从万维网上选择下载原始文档,这部分工作一般由多个具有一定启发策略的机器人完成。为提高机器人的工作效率,可以先根据目标向资源查询站点(如 Yahoo,AItavista 等)发送查询请求,根据返回的查询结果选择采集站点,再运行多个机器人程序,根据一定的启发策略并行地对结果集中的站点进行遍历,采集原始信息。特征匹配是利用挖掘目标特征判断源信息的相似度,即进行相关信息的提取。

网上的信息源各式各样。网上信息挖掘机拥有查询模式匹配规则的知识。在网上提取信息时,它根据用户的信息需求先把内部的查询模板转换成各种信息源能理解的相应查询模式,譬如,SQL 查询模式被用于关系数据库信息源。它还必须拥有本地信息库的知识,以避免信息的重复搜集。它要能定期监视网上的信息源,将新发现的信息通知给用户。

2.2.3　分类标引器

分类标引器是一个自动分类器,它以现在常用的分类法(如《中图法》、《科图法》等)为主要依据,借助语言库中的信息,对收集到的页面信息进行自动分类。分类标引器的具体运作是:对收集到的信息进行元数据(metadata,包括页面标题、作者、摘要、关键词、文件大小、是否有图像、最新修改日期等)抽取,建立索引,并根据这些元数据信息对 URL 进行分类标引。元数据的抽取和 URL 页面的分类,可以自动进行,也可以人工辅助进行。

2.2.4　人工辅助系统

包括信息采集和分类标引的人工辅助。本系统提供一个界面良好的人工辅助信息采集程序,它能够连接到 URL 上,辅助提取 URL 页面信息,进行页面采集和信息分类标引的人工辅助工作。其特点为:它同时具有浏览器功能和数据存取功能,能够对某一页面的元数据进行抽取,并存入数据库;对页面进行分析,显示这一页面下 n 层链接的基本情况;对某一个站点进行自动爬行,并对搜索到的页面进行元数据抽取,存入数据库。

2.3　信息映射系统

该系统负责建立起用户和信息的对应关系。系统通过计算机找到它们的对应关系后,首先要确认用户准备使用什么工具(如 PC 机上的 IE、字符终端等)及采用什么方式获取所需信息,用户接收各种信息的最佳时间和方式以及所能接收信息的容量等;然后根据这些情况,在适当的时间将适当的信息主动"推"到用户的桌面上。

该系统给权限让用户可以选择自己所需要的信息,从而不仅为用户提供方法来修改和扩充信息分类知识,而且还可以从因特网上其他检索系统中学习这些知识。

2.4　信息安全系统

网上主动信息服务系统是基于 Browser/ Server 的 3 层构架,安全问题主要集中在服务器端及信息的传输过程。信息安全子系统按以下方式建立:

首先,紧贴系统服务器所在地的局域网外部建立一道防火墙,在开放与封闭的界面上构筑保护层,防止内部的重要信息泄露给外部的非法用户,同时阻止外部有害信息进入内部。防火墙是多种信息安全技术手段的综合利用,如身份验证技术。

其次,是进行数据加密,即对系统数据库的保护。加密主要是通过某种算法,将数据变换成只有经过密钥解密后才可读的密码来加以保护,使未经授权的非法访问即使得到了数据也无法解读它们。

第三,容错技术。这是为防止网络某个元件的失效而设计的备用方案。主要包括容错用户、硬件容错、容错存储、容错电源、容错网络等公匙加密技术,用来防止信息代码在网络传送过程中被篡改。

3　实现网上主动信息服务系统的技术关键

3.1　网上信息挖掘

网上信息挖掘中最大的技术难点是对除文本以外的多媒体信息的挖掘。其首要问题是要制定多媒

体信息的表述规范,形成国际标准和通用工业标准,这有待于全世界信息业界人士的共同努力。此外,网上信息挖掘的技术关键还存在于以下3个方面。

3.1.1　目标样本的确立

确定用于进行特征提取的目标样本是网上信息挖掘的关键一步。只有选取了适当的目标样本,才能保证信息挖掘的成功。

在主动信息服务系统中,我们遵循"用户中心"原则,由用户选择确定挖掘目标的样本。这就使得对网上信息挖掘机的归纳学习、机器学习和统计分析等人工智能化的程度要求很高,它要在与用户有限的信息交互过程中准确迅速地选择适当的目标样本。就目前人工智能的研究状况看,还有大量的工作要做。

3.1.2　各种词典的建立

网上信息挖掘机的智能不断完善是以词典的完善为基础的。建立完善的用于特征提取和词频统计的主题词典、同义词词典、近义词词典、蕴含词词典等各种词典,并使之最优化,是一项长期、细致的工作。特别是东方文字的词典的研究,目前在世界上还处于起步阶段。

3.1.3　其他相关技术

如特征提取中的权值的计算方法、根据测试样本进行反馈调整时的调整依据和标准、信息采集机器人的启发策略的制定、信息与用户需求相映射的映射规则、机器人怎样在动态网页上获取有用信息、机器人怎样处理收费网页、怎样采用多线程和多进程相结合的机制提高机器人程序的效率等问题。

3.2　**中文信息处理**

因本系统立足于国内,故对中文信息的处理是一个关键问题。特在此提出。

当网上信息挖掘机获取了一个包括中文的HTML文件后,主动信息服务系统就面临着中文信息处理的一系列问题。例如,如何提取其中的概要信息(包括标题、作者、时间、描述、副标题、关键词、URL)以及站点类别等;如何实现对中文的国标码和Big5码或其他的中文常用码的识别、转换等。

3.2.1　中文词的自动切分

中文句子中比较简单有效的分词方法是基于词典的机器分词法,这需要建立大型的切分词库,如1995年全国智能接口评测大纲就选用了《现代汉语词典》和《汉语拼音正词法》中的词作为切词词典。

由于通用切词词典包含了大量的不会成为特征项的常用词汇,为了提高系统运行效率,可根据挖掘目标设立专用切词词典,而舍弃庞大的通用切词词典,这样可以在保证特征提取准确性的前提下,大幅度提高挖掘系统的运行效率。在进行词频统计时,还应考虑到自然语言的多样性,建立并使用相应的同义词词典、蕴含词词典等辅助词典,以提高挖掘的准确度。

3.2.1　中文信息自动标引

网页信息标引是指从网页文档中提取出一组能最大程度上概括其内容特征、可作为用户检索入口的关键性信息,用该组信息对文档进行标引,使用户可以通过输入关键信息检索到该文档的简要信息,进一步点击可查询到该文档。

中文自动标引是至今没有完全解决的问题,实际已有的应用多限制在某个特定领域,如金融、法律、化学等,目前以关键词和全文标引为主。全文标引方法主要问题是信息快速膨胀问题,关键词标引则对专指词(如地名、人名、品牌名和专业词等)无法准确表达。因特网中文信息自动标引主要应解决的问题有:如何提取准确反映网页概念的内容,用尽量小信息量标引巨大的网络信息内容(巨大的信息量将导致标引信息数据库快速膨胀),并且便于快速检索入口;研究一种以概念检索为主又能适应特定专指词检索的标引新方法。

3.2.3　中文信息自动分类

信息自动分类是指建立并维护一套完整的分类目录体系,根据文档的信息特征,计算出与其相关程度最大的一个或多个分类,将文档划归到这些分类中去,使用户可以通过浏览分类体系直接查询到该文档。

中文信息自动分类应解决的主要问题:首先建立反映因特网信息内容,适合广大公众查询需求的信息分类体系;产生和信息分类体系相对应的分类词典(分类词典由分类体系、类相关的词,及反映相关程度的相关系数组成);寻找文本信息归类算法模型。文本归属于某个类的归属度计算,精确完成自动归类处理。目前简单化归类算法对归类误差太大。分类词典应具有新词和相关系数自动维护功能,克服由于词典收词不全面对归类准确性的影响。

3.2.4　中文信息自动检索

用户通过输入由关键信息组成的查询条件、选择查询所限的分类范围、指定查询的限定集合,组成提交检索需求的表达式;系统从标引和分类结果的指引库中找到满足用户要求的文档集合,按照文档与查询要求的相关程度、从大到小的排序和一定的格式输出

到用户界面。

目前,因特网主要站点每年标引的网页为 2000 万～4000 万,几年内网页总量将达数亿,标引信息存储量达百 GB。在这样的海量数据中要实现秒级检索,需要解决以下问题:把标引信息、分类信息和原始简要信息构建成一个适应快速检索的数据结构;规划设计各种检索需求的表达式,检索表达式实际执行的转换及二次检索;建立同义词和中英文翻译词典扩展检索结果;对满足检索条件的所有信息按某个原则进行排序等。

参考文献

1 张灵玲等. 基于 Internet 的课件信息发现和收集 Agent 的研究. 计算机研究与发展,1999(4)
2 卢朝晖. 一种 Internet 自动信息查询模式. 微型机与应用,1998(5)
3 张晓辉等. WWW 上的信息发现与搜索引擎技术. 小型微型计算机系统,1998(6)
4 Styczynski J. Make a Web site sing with free push technology. Information Outlook, July 1999
5 Mitchell F. Wyle. Effective Dissemination of WAN Information. http://www.vhdl.org/%7Ewyle/diss.html

注 本文是国家社科基金(00BTQ010)研究成果之一。

焦玉英 教授,博士生导师。通讯地址:武汉大学信息管理学院。邮编 430072。
晏 凌 武汉大学信息管理学院 98 级硕士研究生。通讯地址:武汉大学图书馆技术部。

(来稿时间:2000-11-06)

(上接第 47 页)设计水平、组织框架及元信息(Meta-information)水平。指南按类组织,有专人负责资源的评论工作,该站点现设有"数字图书馆员奖"(Digital Librarian's Award),按月评出当月最好的指南。

4.2 因特网信息资源评价工具的使用

上述因特网信息资源评价工具可以帮助图书馆及其用户快捷、有效地评价与选择因特网信息资源,应充分加以利用。需要指出的是,由于评价工具本身存在的差异,图书馆及其读者应熟悉、了解各种因特网信息资源指南与评价网站的优缺点,扬长避短。

印刷型因特网信息资源评论工具存在信息滞后问题,使用时应注意出版时间,并上网验证。

因特网信息资源评论网站对因特网信息资源虽然有一定的评价、选择标准,但在标准的掌握尺度、侧重点上是有所区别的。Megellan 在评价因特网信息资源时,特别强调"吸引力",即网站是否有创新性,对眼及耳是否具吸引力?是否有趣?是否是热点、内行或酷等?

有些著名网站,对网站资源的评价是建立在某种带主观与偏见概念基础上的。如某网站评价因特网信息资源的依据是概念"酷"(Coolness)而非信息内容,其出版者声称"cool 即是选择时我的意见与见解"。

一些网站单纯借助某一技术自动查寻因特网信息资源,其结果可想而知。

由于设置目的与使用对象不同,因特网信息资源评估网站既有面向各专业人员的专业资源评价与指南,又有面向一般公众的流行网站资源评价与指南,对图书馆及其用户,前者更有价值。

某些因特网信息资源指南,多是相关网站的罗列,因此所列出的因特网信息资源的质量及完整性无保证。如 Yahoo,请求各网站给予 URLS,再对各种网站资源进行分类并将其加入到数据库中,既不能保证所选网站的质量,也不能保证无遗漏。

参考文献

1 T. Matthew Ciolek. Today's WWW-Tomorrow's MMM? The Specter of Multi-Media Mediocrity. Computer 29, (January) 1996:106～108
2 James Testa. Current Web Contents: Developing Web Site Selection Criteria. http://www.isinet.com/hot/essays/23.html
3 NISS. Net First Collection Development Policy. http://net-first.ac.uk/mission.html
4 Alastair Smith. Testing the Surf: Criteria for Evaluating Information Resources. The Public-Access Computer Systems Review 8, 1997(3)
5 蒋颖. 因特网学术资源评价:标准与方法. 图书情报工作,1998(1)
6 孙兰,李刚. 试论网络信息资源评价. 图书馆建设,1999(4)

罗春荣 中山大学图书馆副馆长,副研究馆员。通讯地址:广州市。邮编 510275。
曹树金 中山大学信息管理系副教授。

(来稿时间:2000-10-16)