

基于 Web 的信息智能获取方法^{*}

彭 岩¹ 涂序彦²

(首都师范大学信息工程学院 北京100037)¹ (北京科技大学信息工程学院 北京100083)²

Study on Web-based Intelligent Information Retrieval

PENG Yan¹ TU Xu-Yan²

(College of Information Engineering, Capital Normal University, Beijing 100037)¹

(College of Information Engineering, Beijing University of Science and Technology, Beijing 100083)²

Abstract The increasing stream of Web information available makes it ever more desirable of network users to retrieval interesting information efficiently. Obviously, AI can be made a good use in the information retrieval area. In this paper, the Intelligent Search Engine, Intelligent Browser, Intelligent Agent and Intelligent Information Push are introduced. Then the related key techniques are presented. At last, a framework of Intelligent Information Push System is discussed.

Keywords AI, Internet, Intelligent information retrieval

1 引言

Web 没有统一的数据模型和查询语言,没有统一的结构,资源分散且信息不断更新,因此用户获取有用信息越来越困难。目前,用户获取信息的方法有两种,即“推送”(Push)与“拉取”(Pull)。“推送”是指信息的提供者主动向特定用户或用户群 Push 特定信息,使用者则通过频道订阅等方式被动地取得信息。这种信息获取方式中信息源的任务较重,它要求信息提供者对用户的需求有准确的了解,否则提供给用户的仍然会有许多无用信息。“拉取”的主要方法是用户进行搜索或浏览。其中,人们使用较多的是通过搜索工具获取信息的方式,但这种方式的弊病在于返回信息过多,随着信息的增多,搜索结果的无效性也会增加。浏览是用户在软件系统分类目录指引下,通过不断链接逐步接近所需要的信息。这种方式的缺点在于,要求用户有一定的领域知识,当用户不知道目标信息的准确分类时,查找将十分困难。为了解决上述问题,将信息获取方式提高到基于知识的层面,提供智能化的信息服务是解决问题的根本和关键。各种 Internet 上的人工智能产品与技术,如智能搜索引擎、智能浏览器、智能体、智能推拉技术^[1]等不断涌现,并正从实验室走出进入商品化阶段。

本文将介绍智能信息拉取及其核心技术,智能体及其所提供的 Web 服务;介绍智能信息推拉技术并设计了一个基于 Agent 的智能推拉系统框架。最后是对本文的总结。

2 智能信息拉取

(1) 智能搜索引擎

搜索引擎是一种可以从各类网络资源中浏览和检索信息的工具。它通过网络流浪者(Web crawlers)、蜘蛛(Spider)、机器人等相关技术,在网上自动抓取和分析网页信息。通常搜索引擎多采用全文检索技术,其核心是关键词匹配。这种方式的缺点是查询结果完全依赖于用户所给出的关键字,而参与匹

配的只是字符本身,并非字符所表达的概念。因此,经常出现答非所问的结果。为了突破关键词检索局限于表面形式的缺陷,人们致力于使搜索引擎具有一定的智能,如基于概念的检索、基于聚类的检索等。

概念检索包括两个内容:一是同义扩展检索,如用户要检索“计算机”,用“电脑”、“微机”可以表达同样的概念。另一个是相关概念的联想检索,如用户键入“手机”,可以联想到一种产品、一种通信工具、性能等等。这些概念关系是从所搜索到的文献中获得的,并且不断在索引过程中进行更新。这样概念检索能够加强系统与人的交互,使搜索引擎具有一定的智能。Yahoo, Altavista, Excite 等提供类似的检索方法,在搜索时不只搜索用户输入的关键字,还可“智能性”地推断用户要查找的相关内容进行搜索。

聚类检索是通过一定的聚类方法,计算出文献与文献之间的相似度,并把相似度较高的文献集中在一起,将文档集合分成若干个簇,要求同一簇内文档内容的相似度尽可能地大;不同簇内的文档相似度尽可能地小。这样,用户只需要考虑那些相关的簇,大大缩小了需要浏览的结果数量。在这样的体系中,主题相近、内容相关的文献便聚在一起,而相异的则被区分开来。当一次聚类生成的簇 S 中仍然包含大量文档时,可以对该簇中的文档再次聚类得到若干子簇($S_{11}, \dots, S_{1j}, \dots, S_{1m}$),直到用户满意。Zamir 和 Etzioni 在文[2]中证明,对检索结果进行联机检索聚类可以显著提高信息检索的效率。

还有一些比较成功的智能搜索引擎如 FSA^[3](Financial Statement Analyzer)和 FAQFinder^[4]等。以 Chicago 大学人工智能实验室开发的 FAQ Finder 为例,它是一个具有问答式界面的智能搜索引擎。在给定一句基于自然语言的英语问句后,它查询 FAQ(Frequency Asked Question)文件以匹配出比较合适的回答^[5]。由于 FAQs 问题成千上万,而且每个问题集中代表一个特定领域的信息,因此,FAQ finder 提供了一个直接而又有效的获取有用信息的方法。FAQ finder 面临

^{*} 本课题研究得到国家863高技术项目(863-511-944-019)的资助。彭 岩 讲师,博士,主要研究方向:计算机网络、人工智能。涂序彦 博士导师,主要研究方向:人工智能及其应用,大系统控制论。

的主要问题是同一问题往往有多种提问方式,因此,FAQ finder 使用了人工智能中的多种方法与技术,如语义知识表达方法^[6],语义网分析及概念匹配技术等,从而找出与用户查询最近似的问题及回答^[7]。

(2) 智能浏览器

智能浏览器是基于机器学习理论设计的智能系统,经过一定的训练后,它可以成为熟练的搜索专家,帮助用户在网络中查找信息。智能浏览器的研究涉及的主要关键技术^[8]包括:特征提取、文档分类、用户兴趣模型的学习与更新及信息过滤等。比较成功的原型系统有 CMU 开发的 WebWatcher 和 MIT 开发的 Letizia。

机器学习是智能浏览器涉及的核心技术,它是浏览器具有智能化功能的基础。以 WebWatcher 为例:WebWatcher 与基于关键字的搜索引擎如 Lycos 和 Altavista 不同,基于关键字的搜索引擎需要用户使用特定的搜索关键字,而 WebWatcher 可以自学习,如:一个“机器学习”的术语可以匹配得到像“神经网络”这样的页面连接,尽管这些词语之间毫无共同点。WebWatcher 是运行在服务器上的系统,它介于用户与 WWW 之间, WebWatcher 伴随用户浏览网络。它将不断地给用户推荐一系列站点并建立超链接。由于 WebWatcher 运行在服务器上,因此它可以记录下数以万计的用户数据来训练自己,从而不断更新知识。如果用户指示某次检索结果是成功的,WebWatcher 会对每一个超链接用代表用户兴趣的关键词加以注释,并存入知识库。

WebWatcher 的任务^[9]主要是在兴趣目标和网页给定的情况下建议适当的页面链接。换句话说,需要如下的目标函数的信息知识:

$$LinkQuality: Page \times Interest \times Link \rightarrow [0, 1]$$

LinkQuality 的值表示“Interest”用户选择“Link”“Page”的可能性,即感兴趣的用户选择链接该网页的可能性。为了不断提高搜索技巧,WebWatcher 采用了以下几种学习方法:

第一种方法是通过对每一个超级链接添加以前用户访问该链接式的兴趣描述完成的,即基于信息检索频度的学习。为了在用户浏览时给出建议的链接,WebWatcher 将当前用户的搜索兴趣与当前页所有链接的描述进行比较。描述与用户兴趣相近的链接,WebWatcher 将会突出显示。向量元素(称为词的权重)使用 TFIDF (Term Frequency Inverted Document Frequency)方法计算。基于这种向量的表示形式,相似度就是两个向量的余弦。

第二种学习方法从超文本的结构中学习,通过给定链接的后续页面中出现的关键词改进该链接的信息描述。这种方法的原理是增强式学习,目的是为了通过网页发现一个能够最大化相关信息量的浏览路径。增强式学习允许主体在特定的环境下学习控制策略以选择最优动作。假设主体通过执行动作从一个状态过渡到另外一种状态。每一个状态 s 下,主体都接收到相应的价值评估。一个动作 a 的质量可以用评估函数 $Q(s, a)$ 表示。 $Q(s, a)$ 为主体在状态 s 下执行动作 a 并选择有动作时可能得到的将来价值评估的折扣总和。如果主体能够学习这个函数,它就可以知道在任意状态下如何动作。准确的公式如下:

$$Q(s_t, a) = \sum_{i=0}^{\infty} \gamma^i \cdot R(s_{t+i+1})$$

其中, s_t 为主体在时刻 t 的状态, γ 则为折扣因子,决定了将来其价值的折扣程度, $0 \leq \gamma < 1$ 。

第三种方法是 Q -学习,即通过用户对已进入页面的价

值评估进行学习。Agent 通过使用 Q 函数选择浏览每一步的推荐链接。运行时,系统将推荐用户兴趣描述所有单词 $Q_\omega(s, a)$ 之和为最高的超级链接。如果搜索兴趣目标与超级链接文字匹配,WebWatcher 将会对其价值给予较高评估。

不同于 WebWatcher, Letizia 是一个运行在客户端的系统,它综合使用了信息挑拣与信息过滤策略收集有关用户浏览习惯的信息,熟悉用户的兴趣爱好,并使用各种启发式策略对现有的知识进行推理,从而实现了一个 Web 上有限资源的智能搜索。

3 智能体 (Intelligent Agent)

智能体是一个具有控制问题求解机理的计算单元。智能 Agent 是将推理与知识表示相结合的能动实体,它在一定条件下能独立运行,与环境发生交互作用,并具有自学习功能,即与智能搜索引擎和智能浏览器相比,用于网络的智能 Agent 具有知识获取与应用的能力、与环境进行通讯的能力及为达到目标而工作的事物处理能力。

目前研究最广泛的是多智能 Agent 系统,它是由一系列网上耦合的 Agent 组成的有机整体,也可由同一环境下的软件 Agent 来构成。智能 Agent 利用机器学习方法,学习并记录用户在查询时的反馈或者用户特征(profile),并利用学习到的知识主动地检索用户感兴趣的信息。由于数据类型以及对数据利用处理方式的不同,在不同的应用系统中 Agent 的结构和组织方式都会发生变化,但对某一特定系统来说,各 Agent 都应当尽可能采用相同的组件和体系结构,以便于 Agent 之间的沟通与协调以及整个系统的整合^[10]。

Agent 设计主要包括以下几个方面的基本内容: Agent 数据类型和数据结构; Agent 信息处理的种类和机制,保证 Agent 之间、Agent 与环境之间进行通畅的信息交流; Agent 行为; Agent 方法; 用于对进入 Agent 的信息进行加工处理或根据要求完成某项任务的功能集合,等等。

网络上的智能 Agent 可以为用户提供外性化和智能化的信息查询服务,帮助用户按照一定的规则处理信息及执行复杂的匹配。如帮助用户搜集同行竞争者 Web 站点中的价格变动信息,提醒用户适时调整价格。用户还可以通过多个智能体来搜集不同的兴趣主题或不同类型的网页。一旦智能体检测出用户兴趣的所在,可以通过电子邮件通知用户,也可以把这些变化增加到其他相关智能体的网页上。由于不同领域用户在查询时对不同信息的兴趣点不同,同一用户在不同时期对一类信息的侧重点均会有所不同。因此,为用户提供个性化的服务显得尤其重要。

目前较成功的智能体系统如:通用智能体 Browser Buddy^[11],它是一个用于组织和链入 Web 页面的基于规则的智能体;一系列用于在 WWW 上寻找特定信息的专用智能 Agent 软件,如:使用信息过滤技术查询电影和音乐的 Fire Fly^[12]、CMU 的 We Doggie^[13](它可以按照用户的兴趣向用户推荐网络资源)和 News Weeder^[14](基于机器学习理论的新闻阅读器)等等。

4 智能信息推送 (Intelligent Push)

信息“推送”(Push),就是通过一定的技术,从信息源中获取用户所需信息,并向用户发送信息的传播过程。它的应用环境不外乎两种:Internet 和 Intranet。而无论是 Internet 还

(下转第154页)

模糊 C-均值聚类,聚类结果可以用来实现水印的强度自适应选择。与一般方法相比,嵌入的水印更具有隐蔽性和不可见性;(2)在频域中进行水印的嵌入,比空域方法有更强的抗攻击能力(比如常见的图像处理方法等);(3)方法思路新颖,容易实现,水印图像具有较好的鲁棒性。

参考文献

- 1 Barni M, Bartolini F, Cappellini V, Piva A. A DCT-domain System for Robust Image Watermarking. *Signal Processing*, 1998, 66(3): 357~372
- 2 Cox I J, Kilian J, Leighton F T, Shamoon T. Secure Spread Spectrum Watermarking for Multimedia. *IEEE Trans. Image Processing*, 1997, (12): 1673~1687
- 3 Cox I J, Linnartz J-P. Some general methods for Tampering with

Watermarks. *IEEE J. Sel. Areas Commun.*, 16(4): 587~593

- 4 Hsu C-T, Wu J-L. DCT-based Watermarking for video. *IEEE Trans. Consumer Electronics*, 1998, 44(1): 206~216
- 5 Huang J, Shi Y Q. Adaptive image watermarking scheme based on visual masking. *IEE Electron. Lett.*, 1998, 34(8): 748~750
- 6 Jain A K, Dubes R C. *Algorithms for Clustering Data*. Prentice Hall, 1988
- 7 Lou D-C, YIN T-L. Adaptive Digital Watermarking Using Fuzzy Clustering Technique. *IEICE trans. fundamentals*, 2001, E84-A(8)
- 8 张伟, 廖晓峰, 吴中福. 一种基于遗传算法的聚类新方法. *计算机科学*, 2002, 29(6)
- 9 张伟. 数字水印中 FCM 算法的应用研究. *计算机科学*, 2002, 29(8)

(上接第102页)

是 Intranet, 用户面对扑面而来的大量信息, 想要及时、准确地获取有用信息并非易事, “Push(推送)技术”^[15]与传统的“Pull(拉取)技术”有明显不同, 其最大的差别是: 用户不再辛苦地花时间上网搜寻, 就可以获取最新的信息, 用户可以预先订阅频道, 设定什么时间、将什么内容送到操作系统桌面上显示出来。图1是我们提出的智能化 push 系统框架。该系统是一个采用开放式 Browser / Server 网络体系结构建立起来的分布式计算机网络, 具有如下特征与功能:

(1) 能自由增加、修改信息频道, 自由选择、优化信息资源。

(2) 系统利用知识库(KB)、方法库(WB)及模型库(MB)和机器学习(ML)、知识发现(KDD)等人工智能的理论和方法, 形成多库协同系统, 提高系统信息获取的智能化水平; 其中知识库用于知识信息的存储和查询。模型库用于模型信息的存储与调用。方法库用于方法信息的存储和调用。在方法库中主要存放决策过程中常用的方法, 如优化方法、预测方法等。

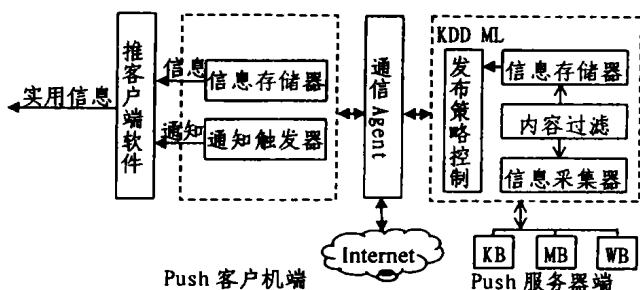


图1 智能 push 系统框架

(3) 具有智能化的信息预采集机制, 同时选择较好的发布策略, 可更加有针对性地向用户推送信息。

总结 WWW 已经成为人们交流和获取信息的重要手段, 随着资源的极大丰富, 广大网络用户在资源使用中所遇到的困难也愈来愈多。因此, 根植于问题求解与知识处理的人工智能技术在 Web 信息获取中大有用武之地, 为各种信息获取

方式的智能化提供理论与技术上的支持。尽管文中所提及的一些原型系统及框架距实用化还有一定距离, 但相信在各方科研人员的共同努力下, 其最终实现为时不远。

参考文献

- 1 涂序彦. 智能信息“推一拉”技术. *计算机世界*, 2000(14)
- 2 Zamir O, Etzioni O. Web Document Clustering: A Feasibility Demonstration. <http://citeseer.nj.nec.com/>
- 3 Mui C, McCarthy W E. FSA: Applying AI Techniques to the Familiarization Phase of Financial Decision Making. *IEEE Expert*, 1987, 2(3): 33~41
- 4 FAQFinder. <http://www1.ics.uci.edu/~burke/faqfinder/>
- 5 Burke R D, et al. Question Answering from Frequently-Asked Question Files: Experiences with the FAQFinder System. The Univ. of Chicago: [Technical Report TR-97-05]
- 6 Cooper E. Improving FAQfinder's Performance: Setting Parameters by Genetic Programming. <http://citeseer.nj.nec.com/16618.html>
- 7 Internet 与 Intranet 中的人工智能技术. <http://member.netease.com/~rygshen/tech/inetai.html>
- 8 应晓敏, 婁文华. 智能 Web 浏览器及其关键技术. *计算机科学*, 2001(9)
- 9 Joachims T, Freitag D, et al. Webwatcher: A Tour Guide for the World Wide. <http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-6/web-agent/www/project-home.html>
- 10 鄢琦, 彭岩, 齐剑锋, 涂序彦. 基于 Agent 结构的 MIS 系统设计与实现. *计算机工程与应用*, 2002
- 11 Canter S. Browser Buddy. <http://www.zdnet.com/pcmag/features/utility/>
- 12 Judge P C. FIREFLY: THE WEB SITE THAT HAS MAD AVE. BUZZING
- 13 Webdoggie, a personalized WWW document filtering system. <http://webhound.www.media.mit.edu/projects/webhound/>
- 14 Lang K. NewsWeeder: Learning to filter netnews. In: *Proc. of the Twelfth Intl. Conf. on Machine Learning*, pp. 331~339
- 15 Web-based Push Technologies. <http://www.itc.virginia.edu/~s/powerpoint/push/tsld006.htm> point/push/tsld006.htm