

信息获取

push 技术

Internet 网

(13)

55-58

计算机科学1999Vol. 26No. 3

# 基于 Push 技术的信息获取方式及其应用

Information Retrieval Mode Based on Push Technology and the Its Application

马红妹 谭庆平 陈火旺

(国防科技大学计算机学院 长沙410073)

TP393

**Abstract** Contrasting with the limitations about the traditional way of information retrieval, this paper analyses the Push Technology and elaborates the high-efficiency of the information retrieval mode based on push technology. Moreover it proposes that applying push technology to the Chinese-Browser so that the performance of the system can be improved. At last it discusses the prospects about the push technology.

**Keywords** Push technology, Information retrieval, Active service, Subscribe, Chinese-Browser

## 1. 引言

随着 Internet 上信息资源的日益丰富和多样化,传统的获取方式已不能满足人们希望快速准确地获取信息的要求。传统的获取方式基于的是 Client/Server 模型,主要采用“Pull(拉取)”技术获取信息。在这种方式下,客户首先向服务器发出请求,然后由服务器返回客户所需的信息,这就存在着某些局限性,为了与服务器建立连接,客户必须确切地知道服务器的地址;由于信息的经常更新,使得用户需不时地访问该站点以获取最新数据。从网络的角度来看,拉取技术基于面向连接的网络协议,提供可靠的数据包传送服务(无错,有序,无丢失),这样虽然简化了信息发布/获取应用程序的设计,但也存在一些不足:网络负载较重时,应用必须承受严重的传输延时;无法避免数据的重复传输,如为将一条信息发送给一千个用户,应用不得不发送一千次。由此可见,传统的获取方式严重影响信息获取的效率。因此,如何使用户能够高效地从无边无际的信海中自动获取与自己有关的优质信息,已成为信息服务行业需要考虑的重要问题。伴随这一需求出现,“Push(推送)”技术应运而生。

Push 技术是 Internet 上一项新兴技术,也是一项很有发展前景的技术。基于 Push 技术的获取方式采用了一种新的服务模型,在这种模型中服务方是主动的,可直接把用户感兴趣的信息推送给用户而无须他们自己来取,从而提高信息获取效率。

本文对 Push 技术及其应用形式进行了分析,并对中文浏览器这一具体应用提出了实现方案,最后文章分析了 Push 技术的发展前景。

## 2. Push 技术

### 2.1 工作原理

从理论上讲,Push(推送)技术是指服务方不需要客户方的请求即可主动地将数据送至客户方,但在实际应用领域中,这种服务的主动性是有一定限度的,即在用户许可的范围内提供主动服务。因此,在信息发布/获取应用中它表现为 Push 服务器自动地搜集用户较感兴趣的信息并将其定期推送给用户。

Push 技术将传统邮递服务中的“订阅”概念引入了 Web 信息传送,形象化地描述了信息获取的过程。其工作流程如下:

第一,用户填写订阅单,该单包括用户个人档案、所感兴趣的信息类型以及要求进行推送的时间等,然后将之提交给信息提供商。

第二,信息提供商按该用户的订阅单收集相关信息并由 Push 服务器推送给用户,客户部件获取信息完毕之后告知用户可读取信息。

### 2.2 特点

(1)无缝连接。指客户部件与 Internet 连接是透明的,即可在无用户交互或最少用户交互的情况下自动建立连接。由于用户与 Internet 的连接方式多种多样,如拨号上网方式,无线 Web 方式或通过局

域网与 Internet 连接等,因此客户部件需要针对不同的连接类型进行相应的连接操作。

(2)灵活的用户设置。用户具有充分的决策权,可设定连接时间、推送内容、本地资源分配等参量。

(3)内容定制文件。用户书写订阅文件,Push 服务器按订阅文件制定传送内容和传送参数,是 Push 技术的重要环节。从用户角度看,内容定制文件使得用户可从信息提供商处选择感兴趣的信息,要求 Push 服务器有选择地推送。从信息提供商角度看,则可依内容定制文件将信息分类以适于不同用户的不同需求。

(4)持久文件传输。指断点重传。当数据传输由于某种原因中断时,可将当前传输状态存于客户部件;当连接恢复时便从断点处继续开始传送。这一特点在传送数据量较大的多媒体信息时非常重要。

(5)有效利用带宽。从客户方看,通过使用空闲时段传送数据,可以达到最大限度地利用带宽。如 BackWeb 公司便采用此技术使用“斯文代理”在网络空闲时从服务器获取文件。从服务器方看,根据组件重用原理将要传送的数据量缩小至最低限度以减少带宽浪费。如 Marimba 公司使用 MD5(Message Digest 5)校验和为文件签名,对不同名字而内容相同的文件便进行重用而不再传送一个冗余副本。

(6)新旧内容自然衔接。更新的内容可以与已有信息相结合。客户部件能确定获取和替换哪些信息,信息的哪些部分;确定将信息存放于何处。

(7)灵活的通告方式。当新的信息到达时,客户部件通知用户可进行读取。依据传送信息的类型和重要性的不同通告具有多种形式,从简单的对话框到具有音频、视频的动画等等。

(8)安全性。能够确保推送给用户的内容是安全的,避免对用户的系统造成破坏。

(9)应用协议。使用 Internet 所基于的 IP 协议组。TCP、UDP 是 IP 协议组中两个最重要的协议。TCP 协议提供可靠的数据传输,但实时性不好;UDP 协议则考虑如何尽快地将数据包发送出去但可靠性差,因此在大量发送数据的同时可能发生包丢失。一些 Push 技术开发商还开发了私有协议,实现特殊服务,如持久文件传输、有效地利用带宽等。但由于私有协议通用性差,因此不易用于有防火墙的网络环境。

### 3. 应用与实现

通过对 Push 技术的工作原理及特征的分析,

可知基于 Push 技术的信息获取方式为用户高效地获取信息提供了有效途径。因此,我们在设计中文浏览器平台时,考虑了将 Push 技术应用于平台设计中从而改善系统性能,进而拓宽机器翻译的应用领域。

按范围和规模大小的不同,Push 技术可分别应用于 Internet 与 Intranet 中。针对不同的应用环境可采取不同的方式进行实现。

#### 3.1 Internet 应用及实现

Push 技术在 Internet 上的应用打破了传统的信息获取方式,减轻了用户上网搜寻的工作量,将个性化的信息直接送至用户,从而提高了用户获取信息的效率。根据对原有系统的继承和扩充程度的不同,Internet 中的 Push 应用一般有三种实现方式:Web 服务器扩展—CGI 方式、客户代理方式和 Push 服务器方式。

(1)Web 服务器扩展—CGI 方式 这种方式使用服务器扩展—CGI(Common Gateway Interface 公共网关接口)来扩充原有 Web 服务器的功能,实现信息推送。CGI 命令可设计出能够对用户输入的信息做出响应的交互式 Web 站点,站点通常把表单(HTML Form)嵌入 Web 页面提供给用户,用户在浏览页面时填写表单进行“订阅”;填写完毕,用户提交表单,由服务器上的 CGI 命令文件对之处理来动态地生成所需的 HTML 页面;最后 Web 服务器将特定信息送至用户。这种方式实质上仍是要用户去“拉”,只不过拉过来的信息具有针对性,是个性化定制的信息,因此是一种最弱意义上的“推送”。这种方式无需特殊的客户端部件,较易构造应用。

(2)客户代理方式 这种方式使用“客户代理(Client Agent)”定期自动地对预定的 Web 站点进行搜索,收集更新信息送回用户。客户代理是对客户端部件的功能扩充。

客户代理对 Web 站点的搜索从其根目录开始直至用户指定的页面,当搜索到该页面后便将所有遍历的内容都返回用户。这就存在一些问题:对 Web 站点的搜索范围较广,返回给用户的内容重点不突出;站点内容的更新与客户代理的自动查询不易同步,不能及时得到更新的内容或重复查找未更新的内容;缺少对站点信息的类型划分,所以需要用户拿到信息后再对之分类以决定取舍;此外,用户还需要控制搜索深度,使用不方便。为克服这些不足,Web 站点需要提供其资源列表和资源的更新状态等信息以配合客户代理的搜索工作,微软公司提出

的CDF(Channel Definition Format 频道定义格式)文件便可用来为Web站点建立结构化的信息内容索引机制以及制订更新计划,指导客户代理的查询。信息提供商发布信息时,不必改动Web站点原有的组织结构,只需建立相应的CDF文件并放于Web服务器上即可。

在这种实现方式中,“主动服务”是由客户代理提供的,因此准确地讲应称之为“智能的拉”。但是从用户角度来看,服务的透明性使得它也可以属于“推送”的范畴,而且很好地继承了原有系统,实现比较简单。

(3)Push服务器方式 这种实现方式对原有系统的改动最大,它提供包括Push服务器、客户部件及开发工具等一整套集成应用环境,经过改动后,这些能够从Internet上向用户推送信息的Web站点被形象化地称为“频道”,用户接收信息就象在收看“专题节目”而且还可以指定其播放时间。

在这里,站点的Push服务器提供主动服务,负责从信息出版商处收集信息形成“频道内容”然后推送给用户,通过它可在信息出版商与用户之间建立起一种透明的多对多通信;专用的客户部件则主要负责接收到的数据及提交指令,并对数据进行处理如存储或显示等。通常Push服务器对信息进行分类组织,可先将新闻条目、文章摘要等信息量较大的数据推送给用户,若用户需要详细了解某方面的信息则可再次获取该专项内容。因此,这种方式减少了传输的数据量,有效地提高了信息获取的效率。但是这种方式对服务方要求比较高,除了进行信息分类和主动推送以外还需要保存和管理用户信息来作为主动服务的依据。与上两种实现方式相比,它是一种“真正的推送”。这种实现方式还可支持私有协议,开发特殊的服务应用。

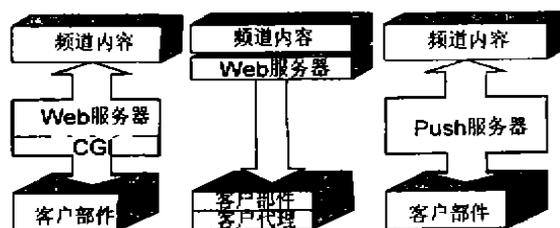


图1 实现方式

### 3.2 Intranet 应用及实现

Intranet(内部网)是一个企业自己的内部网络,它同样采用了基于Internet的技术,如E-mail,Web

浏览和文件传输等,因此Push技术也被引入了Intranet并开发出多种应用。企业网中Push应用的实现方式大致有两类:一是使用内部服务代理代替网关;二是利用一种比HTTP更有效的私有协议,使用私有协议实现的Push应用可改善网络传输性能,提供实时的特殊服务。与之相比使用服务代理的方案则比较简单,较易实现。这种方式使用专用的服务器代理定期地从Internet上获取信息并存储起来,然后再由它把这些信息送给要求它的内部客户机。这样便可有效地将若干客户均通过网关要求信息的多次访问转化为一次访问、多次使用,从而减少了用户的查找时间和上网费用,提高了信息获取效率,另一方面也可有效地解决网络安全问题。

### 3.3 特定应用

这里的特定应用即指将Push技术应用于中文浏览器平台。

由于Internet上的大部分信息是英文信息,对于英文水平一般的大众用户来说,要想浏览并发掘出有价值的信息则存在着一定的困难。因此,我们研制开发了具有英汉翻译功能的中文浏览器平台,它为用户提供一个全中文的浏览环境,使得广大用户能够方便地使用Internet信息。中文浏览器虽然解决了语言方面的问题,但它仍是由用户去“拉”信息,并没有摆脱传统的信息获取方式存在的局限性。Push技术的出现给我们带来了一个很好的思路:将其应用于中文浏览器平台,为用户提供主动服务,改变用户获取信息的传统方式,从而提高用户获取信息的效率,改善系统的性能。

为用户提供翻译和推送服务可采用两种不同的实现方式:一是代理服务器方式,它首先由代理服务器获取英文信息,经翻译后再推送给用户。国内的ISP可以采用这种方式来为用户提供“翻译-推送”服务。另一是对客户端部件(一般指浏览器)进行功能扩充,由它提供翻译服务和推送(实际为智能拉)服务。这种方式对使用环境无特殊要求,适应性强;而且目前许多浏览器不具有这些功能,因此我们设计了这一“智能中文浏览器”,它是前面所述Internet应用实现方式二的特例。

“智能中文浏览器”主要由界面管理、浏览、信息搜集和翻译等功能模块组成。其中,用户界面管理为主控模块,根据用户输入形成请求送至相应处理模块;浏览模块处理用户通常的浏览请求,提供Web服务;信息收集模块处理用户的预定请求,对预定的站点进行搜索并收集发生更新的信息送回用户;翻

译模块将获取的英文信息翻译为中文,是该系统的关键环节。除此之外还有通信联网、CACHE 管理以及本地磁盘管理等模块来改善系统的传输性能并对脱机浏览提供支持。

“智能中文浏览器”的基本结构划分如下图所示:

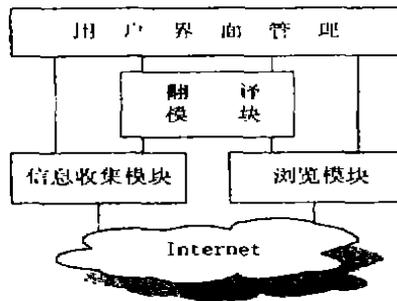


图2 结构划分

#### 4. 发展前景

Push 技术为信息资源的获取提出了更为合理的服务模型,但其发展还不十分成熟,存在着一些问题,如“信息垃圾”问题和带宽浪费问题。因此,寻找一种新的网络通信技术来改善这种现状是 Push 技术进一步发展的首要任务。

多目广播(IP Multicast)是九十年代新推出的网络通信技术,它提供了一种优于传统 Client/Server 模型的通信方式。多目广播可以在网络上为要求大量带宽的应用提供新的途径,并将从根本上改变网络的体系结构。

多目广播以一种优化的方式进行多点到多点间的数据交换。无论有多少用户,数据的发送仅有一

次,必要时在网络上进行数据复制从而减少冗余的网络流量,因此,多目广播提高了网络接口吞吐率和网络资源利用率,为发送大量数据包提供了一种有效的方法。

多目广播与单点传送相比具有很多优点,如:1)节省了传送相同数据量所需的网络带宽;2)允许完全被动的和透明的通信,即接受者可以不知道谁是服务器,服务器也不用考虑谁是信息的接受者;3)易于扩充为大规模的用户组,如欲接收同一条信息的一千个用户。

基于多目广播的 Push 技术将原有的多个单点传送转变为多目广播,利用多目广播的优点解决了在主动服务过程中造成的大量带宽占用问题,在支持多目广播的物理网络中 Push 服务器才可充分发挥优势。因此,可靠的多目广播技术为 Push 应用带来了一个发展的新契机。目前,多目广播技术仍处于研究阶段。如何有效地发展 Push 技术,提高信息获取的效率将是一个新的挑战。

#### 参考文献

- 1 Buchwitz L. Monitoring Competitive Intelligence Using Internet Push Technology. Competitive Intelligence Review, 1997
- 2 Push Technology hybrid application cookbooks, Intel Corporation, 1997
- 3 Andrews W. Planning for Push. Internet World, May 1997
- 4 Davey T. Push Meet Search. Information Week, 1997 4
- 5 Johnson V, Johnson M. How IP Multicast Works. IP Multicast Initiative(IPMI)

(上接第47页)

- IEEE Trans CAS, 1988, 35(10):1257~1272
- 3 Chua L O, Yang L. Cellular neural networks: Application. IEEE Trans. CAS, 1988, 35(10):1273~1290
  - 4 Chua L O, Roska T. The CNN paradigm. IEEE Trans. CAS, 1993, 40(1):147~156
  - 5 Roska T, Wolf D, Kozek T, Terzjaff R, Chua L O. Solving partial differential equations by CNN. In: Proc. 11th Eur. Conf. on Circuit Theory and Design. Davos, 1993. 1477~1482
  - 6 Gobowic D, Xaghloul M E. Analog cellular network with application to partial differential equations. In:

- Proc. IEEE. Int. Symp. on Circuits and Systems, 1994 359~367
- 7 Kozek T, Roska T. A double time scale CNN for solving two-dimensional Navier-Stokes equations. International Journal of Circuit Theory and Application, 1996, 24:49~55
  - 8 Henseler J, Membran P J B. A Cellular Neural Network Model Based on a Vibrating Membrane. International Journal of Circuit Theory and Application, 1992, 20:483~496