一种 WWW 上虚拟信息网络模型*)

A Virtual Information Network Model for WWW

杜晓晨　詹志远　梅卫锋　张　昱　徐永森

(软件新技术国家重点实验室　南京大学计算机科学与技术系　南京 210093)

Abstract　The Information Retrieval(IR)technology for WWW is becoming the hotspot of network research nowadays. According to the limitations of prevailing IR models based on the Client/Server architecture, we put forward a novel IR model for WWW-Virtual Information Network(VIN) model in this paper. The working process of the model, the construction of the VIN and some other key technologies in implementation are discussed while an Active Information Discovery By Content(AIDBC)system built upon VIN is also introduced.

Keywords　Information retrieval, Virtua information network, Virtual site

1　引言

国内外对于 WWW 上的信息发现技术进行了不少探索和研究，开发了一些实用系统，其中应用广泛的有 Yahoo!、AltaVista 等。这些实用系统本身大多基于客户机/服务器体系结构。使用中人们发现采用这种结构的系统普遍存在以下几方面的缺陷：

1)脆弱性：用户必须先登录到相应的远程搜索服务器上才能进行基于远程服务器的信息搜索。一旦远程搜索服务器出现故障或不可连通，用户就无法进行信息的搜索工作。

2)滞后性：用户的搜索请求通常在服务器本地事先生成的数据库中进行处理。搜索服务器通过定期收集各地信息资源列表完成本地数据库的更新，相对滞后于实际信息数据的更新，使得用户得到的往往是过时的信息列表。

3)无导向性：用户无法通过搜索服务器返回的结果判断信息资源当前的状态，也无法知道链接到相应信息资源所需要的响应时间。因此，在不少情况下，运用这些系统并不能给用户以明确的连接导向，用户的信息发现过程仍然带有相当大的随意性。

4)武断性：信息发现服务的提供者强迫用户接受他提供的信息分类方法、关键字的理解、用户界面，而不顾用户的文化背景，传统习惯等。

5)无隐私性：采用这种方式进行信息发现，用户的查询活动、研究兴趣暴露于信息发现服务提供商面前，缺乏对隐私权的保护。

这些问题主要是由于这些系统采用的客户机/服务器的体系结构而导致。在客户机/服务器体系结构下，用户和信息发现服务的提供商通过 Internet 连接，受网络环境的限制，服务器端不能采用实时的下载查询页面的方法，必须在服务器端维护海量数据库，数据库的更新不及时就会导致滞后性。客户机端对服务器端是透明的，服务器端无法知道用户所在的网络物理位置，无法知道用户和目标信息页面的网络连接状况，因此无法给出连接的指导。信息发现服务的提供商只能用自己的理解来处理来自方方面面的用户需求，这种理解对用户个体来说常常是不合要求的。

针对采用客户/服务器结构的信息发现系统存在的问题，我们提出了一种新颖的 WWW 上的信息发现模型——虚拟信息网络。本文提出了虚拟信息网络(VIN)模型并与现有模型作了对比，探讨了 VIN 模型的若干的实现技术并介绍了我们基于 VIN 模型设计的主动式信息发现系统 AIDBC。

2　以用户为中心的完全分布式模型——虚拟信息网络

2.1　虚拟信息网络的定义

就信息发现而言，WWW 上信息分为两大类：一类是数据信息本身，这是用户真正感兴趣的；另一类是元信息，用来描述和表征第一类信息的属性特征。

对于数据信息，根据其所属领域的不同和各个用

*)本课题为国家自然科学基金和江苏省应用基础计划资助项目。杜晓晨　硕士研究生，主要研究方向为 Internet 上的信息发现技术和面向对象软件分析测试，詹志远　博士，美国华盛顿州立大学，主要研究方向为 Intermet 上的信息发现技术，梅卫锋　硕士研究生，主要研究方向为 Internet 上的信息发现技术，张　昱　主要研究方向为 Internet 上的信息发现技术，徐永森　教授，主要研究方向为软件工程和软件理论。

户的爱好，人们又常常将其分为若干类别。针对特定的用户，他所关心的所有数据信息的集合可以表示为形如(class,key-list)的二元组的集合。其中 class 代表信息类别；key-list 代表关键字列表，其长度与信息分类标准有关，由对应的信息类别的粒度所决定。

对某个用户而言，针对某种特定的信息(class,key-list)，他真正感兴趣的页面是所有物理信息页面集合中的一个极小子集 A，

定义 1 (A,(class,key-list))称为一个虚拟站点 VS。

定义 2 (OVS,VS)为一条虚拟路径 VP。OVS 是用户所在的网络结点。

定义 3 ({VS1…VSn},{VP1…VPm})为一个虚拟信息网络 VIN。

虚拟信息网络具有可伸缩性和不确定性。可伸缩性是指随着用户兴趣的不断变化，用户的虚拟信息网络的构成也相应地改变，可以加入和删除相应的虚拟站点。不确定性是指由于网络系统本身的不确定性带来的虚拟信息网络的物理构成站点中的某些站点可能处于事实的不可访问的状态。

2.2 虚拟信息网络的工作流程

新模型从用户的视角看待 Internet，认为 Internet 是与特定信息相联系的虚拟站点的集合，每个页面仅仅在用户需要的情况下才成为特定信息的组成单元。用户为自己提供信息发现的服务，既是信息资源的检索者也是信息资源的使用者。其工作流程可以表示为：

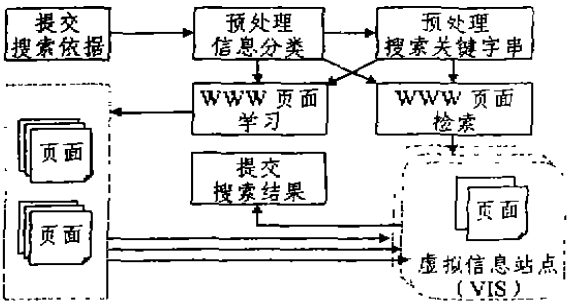

图 1 VIN 的工作流程

用户利用系统获得服务的过程通常也分为三个步骤：

1)用户提交搜索依据：搜索依据包含所要求的信息类别以及特征字串。

2)系统搜索虚拟站点：根据用户提交的信息类别找到对应的虚拟站点，连接虚拟站点的物理组成页面，运用特征字串作为搜索的关键字进行匹配，按照一定的匹配度返回页面信息。

3)系统提交信息：将满足一定匹配度的信息提交给用户。

VIN 模型的特点在于

1)用户可定制的搜索范围。实际上就是相应信息所对应的虚拟站点，虚拟站点的建立过程就是用户定制搜索范围的过程。有关虚拟站点建立的过程详见 2.3 节。

2)完全分布式的体系结构。如前文所述，客户/服务器结构不适合建立用户可定制的搜索范围机制，完全分布式的结构则不然，因为在完全分布式的系统中，信息发现系统实际上存在于用户本地，对于一个用户而言，其感兴趣的信息类别和站点均十分有限，就算使用数据库进行管理规模也不大，不用耗费大量资源。而且，由于用户自己的信息被保存在本地，不存在网络传输和远地保存可能带来的隐私暴露的问题。

VIN 模型能够较好地解决当前 WWW 上信息发现存在的一些问题。

2.3 虚拟信息网络的构建

虚拟信息网络构建的核心是 VIS 的构建。每当用户第一次提交一种新的信息分类的搜索请求时，系统就开始一个对应的 VIS 的构建过程。VIS 的构建流程大致如下：

1)用户提交(class,key-list)和初始 URL 列表(url1,url2,…,urlm)。

2)分析用户提交的信息分类 class，初始化一个初始 VIS。

3)根据用户提交的(url1,url2,…,urlm)连接相应的页面，如果响应时间大于用户设定值则认为不可连通。如果所有页面均不可连通，转向 8)，按照以下步骤处理可以连通的页面。

4)进行页面的自动学习过程：将正在学习的页面的 URL 记录下来，分析目标信息页面，把页面中的数据信息与元信息提取出，从指定的模式库中匹配模式，构造模式实例，模式实例中包含了文本信息串、该模式的类型、模式在信息查询中的权重、信息查询策略等信息，把页面中的所有超文本链接插入到后备 URL 列表(url'1,url'2,…,url'k)中。

5)进行信息查询：根据用户提交的 key-list 查找字典，得到一系列领域等价的关键字列表；运用关键字匹配算法对模式实例树库中的模式实例树进行分析，如果匹配度大于设定值则认为匹配满足，将数据信息返回给用户，同时把该 URL 加入虚拟站点。

6)当一个页面学习和查询完成后，优先学习用户提交的初始 URL 列表(url1,url2,…,urlm)中的下一个页面。

7)当初始 URL 列表中的所有页面均学习完成后，根据用户的要求决定是否开始自主的页面学习。如果用户要求，系统将根据一定的优先级策略选择后备 URL 列表中保存下来的超文本链接，判断是否属于已经学习过的页面，如果已学习则取下一个连接否则进

行学习，学习的方法同3)。如果用户没有要求，转向8)。

8)系统的自主学习过程将终止于用户事先设定的条件，如匹配数或学习时间。

9)询问用户是否有其它的URL列表提供学习，若否，学习过程结束，完成虚拟站点的构建；若是，则转向2)。

10)返回构建失败信息，删除初始化的虚拟站点。

在上述流程中，有几个比较重要的实现技术需要特别说明：

1)学习过程的自动化、智能化。一个好信息页面的自动学习系统应该是自适应的，能够自动学习不同站点的特征，并相应地调整模式实例树的生成策略；应该是可以配置的，允许用户配置具有自己个性的信息发现的范围，模式实例树生成策略；应该是自学习的，根据本次学习的成果自动地修改扩充知识库，使工具越来越友善，越来越适应于主人的习惯。因此，好的学习模块的实现需要用到机器学习的方法。

2)领域相关的关键字列表字典。虚拟站点的构建目的不仅仅是为了满足用户的本次搜索请求，更重要的是为了用户在今后搜索本领域内的信息时提供一个相对稳定的搜索范围。因此，在3)的页面学习过程中仅仅依靠用户提供的关键字串进行匹配操作来判断该页面是否应当加入虚拟站点是远远不够的。一种比较好的实现策略是建立一个关键字列表的字典，该字典按照各个信息分类进行组织，将一个关键字与本领域中等价的若干关键字组成的列表对应起来，扩大了页面学习时匹配的范围，为建立一个领域信息相关的虚拟站点提供了可能。这个字典本身是用户可以修改维护的。

3)超文本链接的时间加权选择。在用户允许的情况下进行自主的页面学习真正体现了信息发现的特点。一个页面选择策略的好坏对于学习的效率和结果有很大的影响。其中最主要的两个因素是页面内链接组织方式以及连接页面的响应时间。在绝大多数站点上，页面内的链接是按照信息粒度从粗到细来进行组织的。也就是说，如果一个页面的信息内容符合用户的需要，这个页面中的超文本链接就有可能比其它页面中的链接更加符合用户的需要，因此它们应当具有较高的学习优先级。

另外，页面的响应时间也不可忽视。在已经学习过的页面中的链接不一定全部指向这个站点上的页面。如果选择了指向响应时间特别慢的其它站点的链接就会在很大程度上影响学习的效率。因此，在指定超文本链接的学习优先级的时候一定要有响应时间的加权。

2.4 虚拟信息网络的自适应调整

虚拟站点是可扩充的，能够根据网络环境的变化作出自适应的调整。所谓网络环境的变化是指网络系统中新出现了url0，它含有虚拟站点VS(A，(class，key-list)中的(class，key-list)，并且存在一个url1∈A，url1的页面中含有指向url0的超文本链接。为了适应这种情况，虚拟站点的构建过程结束后，可以保留后备URL列表。当获取信息时，系统将页面中的超文本链接与后备URL列表中URL进行比较，如果是尚未学习过的新链接，则进行新页面的学习，将学习的结果加入虚拟站点。

3 基于VIN的面向信息内容的主动式信息发现系统AIDBC

基于以上提出的虚拟信息网络的模型，我们开发了面向信息内容的主动式的信息发现系统AIDBC(Active Information Discovery By Content)。AIDBC采用完全分布式的体系结构，以用户为中心构做虚拟信息网络，每个用户的信息发现工作组成虚拟信息网络中的站点。AIDBC充分尊重用户的个性，用户可以自主地、自治地、自由地进行信息发现的工作；信息发现的结果不再是传统的松散凌乱的信息集合，不同的信息发现的主体可以组织成一个"社会"，"社会"中的成员自由地共享他人信息发现的成果。AIDBC的信息页面学习、信息查询具有一定的智能，具有一定的自适应性、可配置性、自学习性。此外AIDBC还维护有反映用户个性的，可扩充的关键字字典。

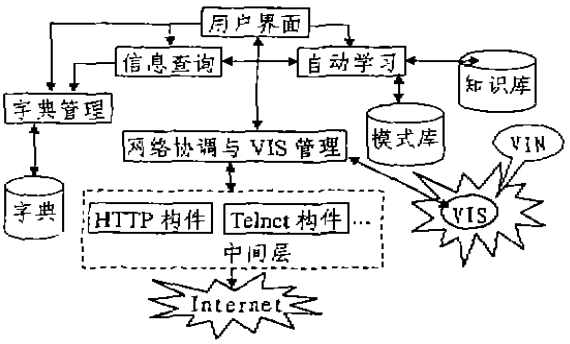

图2 AIDBC系统结构图

AIDBC系统主要由用户界面、网络协调与VIS管理、字典管理、自动学习、信息查询等主要模块组成。

用户界面模块：负责与用户交互，提供多种风格供用户选择，提供配置个性化系统的工具。

自动学习模块：分析网络协调与VIS管理模块取回的目标信息页面，运用知识库中的知识对其进行信息的特征化和模式匹配工作，生成供信息查询模块进行信息查询的模式实例树。模式定义为各种信息页面的特征结构、信息存放区格式，以及供信息查询的各种权重、查询策略的描述，模式实际上是模式实例的模板，模式和模式之间可以有继承，耦合关系。模式实例

是模式的例化，是目标信息页面的一定的标记结构以及其中的文本信息匹配了模式产生的。自动学习模块中的模式的设计对于整个系统的智能起关键的作用。一个理想的学习模块将能够根据知识库中的知识，对目标信息页面进行分析来获取其特征，并将其应用于一定的模式，存储成特征化的信息的形式，以备查询模块使用。各种特征化的信息组织成树状的结构，叫做模式实例树。模式实例树完全反映了信息查询需要的信息。学习模块遇到模式库中没有的模式会自动地记录下来，当同样出现的模式达到一定频率，自动将新模式添加到模式库中。

信息查询模块：从字典管理模块获得同义关键字用以对模式实例树进行关键字匹配。模式实例树是对目标信息页面以链表的形式的结构化重组，模式实例之间可以嵌套。在模式实例树中各种结构的匹配的权重和查询策略是不同的，例如，在信息页面的提供者看来，标题信息的重要程度应该比一般的正文大。因此，应该具有比较高的权重。

字典管理：字典管理的主要功能是提供维护、查询和扩充字典的操作。管理从一个关键字映射到一组同义关键字，信息查询是使用这些同义关键字。对同义的理解，每个用户都可能有差别，因此字典管理除提供智能扩充同义关键字表的功能外，还要提供用户配置关键字表，选择同义关键字的功能。同义关键字在字典里是以集合组织的，同一集合中的关键字互为同义。不同集合之间的可以有不同的关键字，多个集合又组成一定的领域。同义关键字的映射一般采用一步映射方法。对于特殊的应用也可采用多步映射，即在第一次映射的关键字中再找映射。

网络协调和 VIS 管理：负责管理访问网络的所有事务，包括获取目标信息页面和共享其他 VIS 的成果。获取目标信息页面是通过中间层调用一定的协议接口，如 http 协议。VIS 成果的共享可以通过 Agent 之间的通信。

中间层：定义了与协议无关的信息检索接口，在相应的构件中得到实现。中间层的设计的主要目标就是在研究目前协议(HTTP 和 Telnet，以及 BBS 系统的特点)的基础之上设计出一个完备的、抽象的、与信息检索和自动学习有关的网络操作集合。整个中间层可能以一个或几个抽象类的形式存在于最终的实现系统中。这部分的设计过程是一个需要反复论证、不断修正的过程，先主要由界面部分和系统管理部分从功能的角度提出操作需求，然后主要由协议构件部分从实现的角度给出修正，最终形成一个更新版本的中间层的设计。

虚拟站点：主要存放用户定义的以虚拟站点为单位成组的网络地址、学习历史、结果文件索引等信息以及网络地址配置、最大等待时间等系统参数。虚拟站点的内容在该虚拟站点初始创建时建立。在用户使用系统的过程中，系统管理将根据界面所记录的用户访问情况对虚拟站点进行维护。

目前，AIDBC 系统在使用中取得了很好的效果，我们用它建立天气信息的虚拟站点，选择的领域关键字为天气、气象、今日天气、天气预报，以国内著名站点的首都在线的首页 http://www.263.net 为唯一初始 URL，在 10 分钟的时间建立起了虚拟站点。该虚拟站点的首页按匹配度组织出十几个有关天气的连接，所涉及的内容有天气预报、天气的新闻等，内容高度地符合了需求，所有的连接都实时可以访问，消除了信息滞后性。在以后的服务期间，虚拟站点会自动有选择地更新。实验表明该系统方便、实用、智能、友好，更重要的是系统的使用很好地反映了人的学习曲线，有着越来越好用的趋势。尚存在的问题是，由于要实时分析学习很多的页面，网络开销大，在网络条件不太理想的情况下，学习的过程可能稍长，这也是保证实时性必然带来的问题。网络开销大的问题，在以后的研究中可以采用移动代理的技术解决，但目前移动代理技术尚不成熟，并且几乎所有的 Web 服务器都不提供移动代理访问功能，开发实用的信息发现工具尚不能采用这项技术。

小结 通过对传统模型的分析可以看出，当前 WWW 上信息发现模型的根本问题在于用户无法定制信息搜索的范围，同时客户/服务器体系结构也从实践上限制了对传统模型的改进，因此，我们提出了一种完全分布式的 WWW 上信息发现模型——虚拟信息网络(VIN)，支持信息搜索范围的用户定制，并且结合我们的课题实践讨论了基于 VIN 构造的面向信息内容的主动式信息发现工具构造中的相关技术。我们相信随着研究的深入，虚拟信息网络模型及其相关技术一定会更加成熟实用。

参 考 文 献

1 Sycara K, Zeng D. Multi-Agent Integration of Information Gathering and Decision Support. In: Proc. of the European Conf. on Artificial Intelligence, 1996

2 Bowman C M, et al. Scalable Internet Resource Discovery: Research Problems and Approaches. Communications of the ACM, 1994(Aug.)

3 Zhong Yutao, Huang Weiyun, Chen Xin, Xu Yongsen, Chen Junliang. A Multi-Agent Cooperation Model in Internet Information Retrieval. In: Proc. of TOOLS Asia'99 & OOT China'99, IAP, 1999

4 Chen Xin, Chen Junliang, Zhong Yutao, Liu Guodong, Xu Yongsen. VIN: A User-oriented and Cooperative Model in Internet Information Retrieval. In: Proc. of ISFST99, ASE, 1999

5 黄伟韵．多 Agent 协同的网上信息发现技术研究：[南京大学硕士研究生毕业论文]．1999

6 詹志远．虚拟信息网络模型及其实现：[南京大学硕士研究生毕业论文]．1999

7 詹志远，黄伟韵，钟昱陶，等．一种主动式网上信息发现模型及其实现．见：第十届全国计算机网络与数据通讯会议，计算机应用．1998