理想的 Web IR 服务模式的研究

刘 悦 冯国臻 程学旗 薄立彦
(中国科学院计算技术研究所软件室 北京100080)

The Ideal Web IR Service Frame

LIU Yue FENG Guo-Zhen CHENG Xue-Qi BO Li-Yan
(Software Division Institute of Computing Technology, Chinese Academy of Sciences Beijing P. R. China 100080)
E-mail: yliu@software.ict.ac.cn

Abstract Web IR presents a new challenge due to the heterogeneity, the dynamic characteristic and the size of the Web. A practical IR system that can satisfy the users' demand is very important, in this paper we research the characteristics of Web IR in detail and give out the ideal Web IR service model: it should include search engine spectrum, search engine hierarchy, search engine cooperative network. We also analyze the key technique of this model, propose a simple way for the Web IR service to deal with the huge-scale of Web resources easily, and test part of the ideas in our prototype system SAInSE.

Keywords Web IR, Search engine, Search engine spectrum, Search engine hierarchy, Search engine cooperative network

1 引言

信息检索(information retrieval, IR)指的是从一个文档集合中查找出包含有与用户的信息需求相关的信息内容的文档或文本。Web IR(Web information retrieval)指的是在 Web 环境下的 IR 问题。Web 的出现，将含量、异构、动态数据的处理这样一个新的课题摆在了人们的面前。本文在分析了 Web IR 的特点，特别是 Web 用户信息检索的特点之后，从用户需求的角度，对架构 Web IR 的理想的服务模型提出了我们的一个构想，并分析了理想 Web IR 模式实现所需要的主要技术支持。

2 Web 的特点

Web 自从80年代后期出现后，一直以惊人的速度发展着，已成为人类历史上承载数据最丰富的信息库。正是由于 Web 本身的固有缺陷，在 Web 上查找所需要的信息却很困难。这就使得 Web IR 比经典 IR 表现出更大的挑战性，广泛地引起了各方面的研究兴趣。Web 的特性我们可以从两上方面加以考虑：一方面是 Web 数据本身的特点；一方面是与用户的检索行为有关的特点。

2.1 Web 数据本身的特点

Web 数据来源广泛，结构性差，而且数量巨大(很多地方以海量来形容 Web 数据)，内容亦是良莠不齐。概括起来有如下特点[1,2]：

·分布式数据 由于 WWW 本身的特点，使得数据分布在不同的地点的计算机和服务器上，而且不同的系统使用的平台可能也不一样，这些承载着数据的计算机通过网络互连，构成了一种物理上的分布式的拓扑结构。由于是通过网络互连，数据的可靠性就不能完全保证。

·数据的不稳定性 由于 Internet 本身是动态的，每天都有新的页面产生和加入，同时也有一些页面和链接被更新，因此 Web 上的数据是变化的。

·数据量巨大 Web 本身以指数级的速度在增长，所以它的数据规模成了一个很棘手的问题。

·结构性差，而且数据冗余 虽然 Web 上的数据是以超文本的形式组织的，但作为超文本页面本身的结构性还是比较差。Web 上重复页面非常多，这就造成了数据的大量冗余。

·数据质量不高 作为一种新的出版媒体，由于大多数情况没有正规的编辑过程，因此常见的印刷和语法错误难免经常出现。

·数据的异构 除了有多媒体数据，各种不同格式类型的文件，还包括各种语言。

2.2 Web IR 的特点

从用户检索行为的角度，长期以来，人们已经认识到了了解和研究用户查找信息的行为方式的重要性，Web 用户的专业背景、教育程度、对 IR 系统的经验程度，用户在 Web 上访问和检索信息的目的有着巨大的差异，但用户的检索行为本身存在着若干规律性和共同的特点。从单个信息检索类型来看，Web 信息查询可以分为如下三类：

·一般信息查询 用户要求的是某一方面的信息，任何属于该方面的页面都是满足用户需求的。例如，查找有关空气动力学方面的文档。

·精确信息查询 用户需要查找某一个具体的文档，其他任何文档都不符合用户的需要。例如，发现某个人的主页，或

*)本论文得到国家973课题资助(课题编号 G1998030413)。刘 悦 博士研究生，研究兴趣：海量信息处理，知识检索与数据挖掘，算法设计与分析，petri 网理论与应用等。冯国臻 博士，研究兴趣：海量信息处理，知识检索与数据挖掘，数据库等。程学旗 博士研究生，副研究员，主要研究领域为：Internet 高性能软件、智能信息处理、知识检索与算法分析、信息安全、计算语言等等。薄立彦 硕士研究生，研究兴趣：人工智能，电子商务等。

者查找某一个指定作者和题目的文章。

·相似信息的查询 这种查询的特点是用户的查询输入是一篇或若干文档(称为种子文档)希望系统找出更多的与所提交文档类似的页面。种子文档可以以 Web URL 或文档全文两种方式提交。

我们认为从用户来看信息检索需求可以分为两类:

·偶然的查询 由于比较偶然的原因查找该方面的信息,对检索质量要求不高。

·经常性的查询 用户的工作或持久爱好所属领域,检索质量要求很高。

在深入地研究了 Web IR 的特点之后,发现万能的搜索引擎不是一个好的策略,我们可以在分类的基础上针对不同的类型采用适合的技术构造搜索引擎,并且各种搜索引擎相互合作,共同提供 Web IR 服务。所以我们提出了理想的 Web IR 服务框架。

3 理想的 Web 服务框架

搜索引擎服务框架的最终目的是为用户提供最优信息检索服务,并且资源消耗最低。通过搜索引擎之间的合作,最大限度地满足用户的各种信息需求。对于泛泛了解的查询请求,整个 Web 规模的综合型搜索引擎能够给出 Web 上该方面最重要的相关信息;而对于经常性、比较深入严肃信息需求的方面,用户能够比较容易地找到该领域的专用搜索引擎,从而得到更加精致智能、更新更及时的信息检索服务。由于 Web 所表现出的巨大差异性,我们认为应该针对不同的服务范围采用适宜的技术构建不同类型的搜索引擎,构成一个 Web IR 的服务框架,该框架的主要要素为:

3.1 搜索引擎的谱系

不同作用范围的搜索引擎构成一个搜索引擎谱系,我们把该谱系中的搜索引擎分为三类:综合型搜索引擎,领域型搜索引擎和资源中心。

在该谱系的一端是综合型搜索引擎,它们的处理对象是整个 Web 上的信息,目前大多数知名的大规模搜索引擎都可以归入此类。它处理 Web 上的全部信息,规模宏大,信息内容无所不包,差异性极大,与之相适应,其策略上应该侧重于:

·全面索引。这一点是综合型搜索引擎的主要优势。索引库大小也是当今各大搜索引擎竞争的一个焦点参数。据估测今天 Web 仅静态公开页面数量已达到2,000,000,000以上,即使采用最先进的技术,目前没有任何搜索引擎能够索引整个 Web。

·有重点地索引。我们认为综合型搜索引擎的目标应该是有重点地掌握 Web 上最重要的内容。

·满足普通查询。综合型搜索引擎服务重点是满足一般型查询。

谱系的另一端是最小型的资源中心,它们针对的是某个非常具体狭窄的主题。它可以是自动检索工具如搜索引擎,也可以是人工维护并提供在线检索的资料库,甚至是静态页面。

位于这两个极端之间的是领域搜索引擎,又称为"垂直搜索引擎"。它们处理的是某种划分标准下 Web 的某子集的信息,它所处理的信息有一个限定的范围,如对限定语种、地域、行业或类型的信息进行处理。

与综合型搜索引擎相比,领域搜索引擎处理的信息由于属于某一领域,有更多的共性和规律性可循,规模更小,面对相对稳定的用户群。在提供检索服务时,应当力求对处理范围的信息本身、用户信息需求、用户检索行为等的特点有更加深入的了解,提供更好质量的检索,通过优质深入的服务抓住对该领域信息有兴趣的一批常规用户群。

我们认为领域型搜索引擎应该在服务的深度而非广度上竞争。具体说来,领域型搜索引擎的服务特点应该为:

·准确。

·智能。

·实时。信息的更新周期更短,对新信息更加敏感。

·其他个性化主动服务。由于针对比较经常和固定的用户,可以通过收集和分析检索历史记录或用户主动定制等手段了解用户兴趣,主动推荐或推送与用户兴趣一致的新内容。

上述的搜索引擎类型的划分是一个定性描述,各种搜索引擎不是截然划分的,其作用范围可能出现重合。

3.2 搜索引擎的层次结构

各种搜索引擎按照其作用范围在人类知识体系层次结构的对应关系,相应地也构成一个层次结构。该结构基本呈现树形,是一个不严格的树形的网状结构,如下图所示。

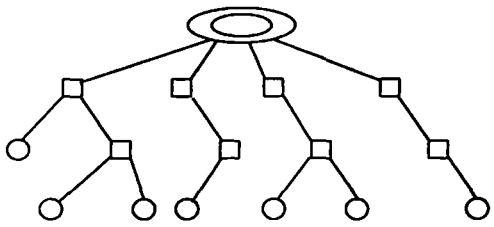


图1

图中的结点代表搜索引擎的类型,根节点(同心环)为综合型搜索引擎,其作用范围是整个知识体系;叶子节点是各个资源中心,中间节点是各个领域型的搜索引擎。双亲节点的作用范围覆盖其子女节点。

这种层次结构框架只是一个概念化示意。在现实中可能有多个搜索引擎对应着图中的一个节点,尤其是根节点,如上所述包括大多数今天的大型综合搜索引擎。另一点是由于领域搜索引擎的划分标准可能不同,非根节点的搜索引擎之间的作用范围可能重合。

3.3 搜索引擎合作网络

由于搜索引擎谱系上从综合到具体不同的位置的搜索引擎共同构成的搜索引擎层次结构形成一个有机的相互合作整体,称为搜索引擎合作网络,而不是一个简单的共同罗列。

合作网络中的搜索引擎的检索结果除了当前常规搜索引擎的检索结果内容之外,如匹配目录、站点、页面,此外特别的还有针对该查询最适合的搜索引擎(可选的),以及来自这些所建议子领域搜索引擎的检查结果。一种可能的检索结果的提交组织方式是采用多帧页面,如图2所示。而当一个搜索引擎对于一个查询没有检索结果时,它自动将该查询转交给自己的上级搜索引擎,直到获得检索结果或到达搜索引擎合作网络的根节点。此时检索结果的提交方式如图3所示。搜索引擎间的自动转发查询请求只在上下级之间进行,即被请求的搜索引擎只向自己的双亲或子女节点上的搜索引擎转发请求。基本原则是尽量向用户推荐有关该查询所在领域的合作网络中底层页节点方向的小范围搜索引擎。仅当查询在所请求的搜索引擎上没有检索结果时,才将用户查询向检索范围更广泛的高层搜索引擎转发。

初始搜索引擎的检索结果	帧1
参考搜索引擎1以后的检索结果	帧2
参考搜索引擎2以后的检索结果	帧3
……	……
参考搜索引擎 n 以后的检索结果	帧 n+1

图2 采用多帧页面方式的一种可能的检索结果的提交组织方式

检索请求自动转发的出发点是将搜索引擎也看作是Web上的一种具有主题属性的资源，尽量将某方面主题的搜索引擎介绍给具有这方面信息需求的用户，使得用户能够得到最适当的信息检索服务，而且对于整个搜索引擎合作网络来说做到最低检索代价，即实现资源的最优化利用。

尝试将搜索引擎资源和用户进行这种匹配的依据是用户信息需求的稳定性。研究表明，一个用户的信息查询请求具有连续性，往往对某领域经常查询，甚至重复完全相同的查询。[Silverstein 1998][7,8]中指出，我们每个人都有自己感兴趣的东西，它们构成我们日常浏览和检索的大部分内容，找到与自己兴趣范围相吻合、建造得比大型综合型搜索引擎更加智能和细致的领域型搜索引擎，对我们的信息查找是很有帮助的。

不能找到合适的查询结果，请参考其他高一级的搜索引擎进行检索。	帧1 帧2
参考搜索引擎1的检索结果。	帧3
参考搜索引擎2的检索结果。	……
……	帧 n+1
参考搜索引擎 n 的检索结果。	

图3 当一个搜索引擎对于一个查询没有检中结果时，其检索结果的提交方式

4 理想 Web IR 服务框架的关键技术

如果用户在返回结果中发现参考搜索引擎 i 的结果最满意，以后再次需要查找这方面信息时便可以直接与该搜索引擎交互。搜索引擎合作服务框架概念的关键点是查询请求对用户透明的适当的自动转发，使得整个 Web 信息检索服务框架在最终用户眼中呈现出整体性和合作性。这就是这个搜索引擎服务框架的核心关键技术所在。成员搜索引擎之间能够透明地转发请求是它们构成一个合作的有机整体的根本要求，这项技术的实现需要以下要素的支持：

·*搜索引擎分类标准*(Ontology tree) 这是实现搜索引擎层次结构的基础，该标准定义搜索引擎有哪些标准类型，各类型之间的上下级关系，该标准是整个搜索引擎层次结构树的规范化描述，标准化地定义了该搜索引擎服务框架的 Ontology 树。即给搜索引擎的分类一个统一的标准。标准 Ontology 树是动态的。随着新的搜索引擎的出现或者原有搜索引擎检索范围的变化，可能会增加新的节点或子树。另外 Ontology 可能会需要子树或节点的修改、删除操作。显然该标准的建立属于 Ontology 范畴，在一个存在巨大异构性的信息库上建立标准 Ontology 体系是非常困难的。在这里需要指出的是，如果搜索引擎的数量是 K 级的，而 Web 站点和页面是 G 数量级的，也就是说搜索引擎比站点和页面在数量级上低了6个级别，搜索引擎的数量是相当友好易协调的。也就是建立搜索引擎的 Ontology 范畴要比对整个 Web 内容建立 ontology 体系容易得多。

·*搜索引擎间交互协议* SSCP(Search engine to Search engine Communication Protocol) 在前面我们已经提到过理想的搜索引擎服务框架中要有一个能够保证检索请求自动转发给不同级别的搜索引擎的机制，所以建立类似 HTTP、SMTP 等，建立在 TCP 连接之上的应用层协议，对于搜索引擎合作网络是必要的，这样才能保证这种理想的搜索引擎服务模式采用请求/应答方式进行交互，用于搜索引擎间、搜索引擎与集中记录之间交换信息。

·*搜索引擎的自标引*(Self Identification) 每个搜索引擎适当地标识自己的检索范围，检索服务特点，并将自己定位于 Ontology 树的某一节点。

标引时的欺骗行为一方面对搜索引擎自身不利：当用户发现该搜索引擎所标识的服务与实际不符合，没有提供自己所需要的信息服务，会离开该搜索引擎，而且留下恶劣映象。另一方面在服务框架中欺骗行为是非常不受欢迎的，一旦发现，作弊搜索引擎会受到警告甚至被其他搜索引擎从合作名单中去除，从而被排除在搜索引擎服务框架之外。

·*集中记录*(central memory) 一个搜索引擎服务框架有一个集中记录有关信息的机制，保存该合作网络的搜索引擎分类标准定义，成员搜索引擎在 Ontology 树上所对应的节点位置，甚至包括成员搜索引擎的描述标引信息。

每个成员搜索引擎都知道该集中记录机制的地址。当自己的检索范围发生变化时，主动向该集中记录报告；定时查询集中记录，获取最新的上下级搜索引擎信息。

·*参与方式*(way of participation) 搜索引擎参与到某个搜索引擎服务框架中有两方面的含义：

1)接受该 Ontology 树分类标准，并在该分类标准的指导下将自己定位于 Ontology 树的某个节点；支持 SSCP，履行约定义务，如向集中记录报告自己检索范围的变化，定期查询集中记录更新上下级信息等。

2)向最终用户提供检索服务时，同时提供推荐适当的搜索引擎服务。当用户的查询请求在本搜索引擎没有检中结果时，向双亲节点上的搜索引擎转发请求，并且只向上一级的搜索引擎转发请求。

对于高层节点处的搜索引擎，重要的是掌握其子树上各搜索引擎情况，并向用户推荐。具体方式可以是直接向用户推荐与其查询高度相关的子树上的搜索引擎；也可以是将查询请求转发到子树搜索引擎，并将它们的检索结果提交给用户，由用户自己决定哪些搜索引擎的检索结果是更加令人感兴趣的。

用户就自己经常性的严肃信息需求找到比较满意的适合的搜索引擎，(领域型)搜索引擎拥有一批比较固定的查询用户，这种用户和搜索引擎之间比较稳定的关系对于用户和搜索引擎双方都是有利的。

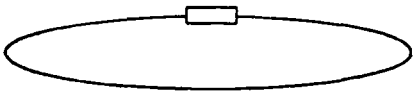
用户可以更多地熟悉并参与领域搜索引擎的运作，例如熟悉其使用规则、高级技巧，了解该搜索引擎在排序中识别的页面结构组织方式(如 Metadata 标准规范)并主动在自己制作的页面中遵守，当自己的页面发生变化时主动向搜索引擎提交，更深入地参与包括向搜索引擎提供有可信度的反馈信息，如就页面、站点的标引或评价提供反馈信息。

随着用户(包括用户制作的页面和站点)与搜索引擎之间关系的稳定，开放公网和受控网、私有网之间的界限逐渐模糊，协调组织更容易进行，Web IR 的一些突出矛盾，如海量规

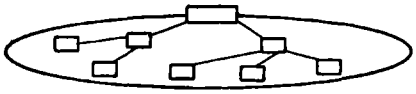
模、异构性、欺骗行为普遍、不易协调等等都会被淡化消减，一些适合更小规模、需要密集利用人工劳动的技术也能够应用，建造更加智能灵活的搜索引擎。

综合采用上述技术，并配合人工/专家劳动和用户参与，在限定领域范围，领域型搜索引擎将提供比大规模综合型搜索引擎更智能、高质量的 IR 服务。一个我们身边的例子是科学院的 FTP 搜索引擎，检索质量可以说是令人满意的。

结论和未来的工作 传统搜索引擎将 Web 看作一个无结构文档库，对每个文档、文档中的每个词一视同仁地进行处理，力图索引尽量多的文档，搜索引擎是用户与 Web 这个巨大文档库交互的唯一接口，该视图可以用图4(a)表示。



(a)传统的搜索引擎与 Web 的关系



(b)在理想的搜索引擎服务框架下的搜索引擎视图

图4 不同视图下搜索引擎与 Web 的关系

而在相互合作的搜索引擎服务框架下，Web 被看作一个分门别类的知识库，如图4(b)所示。综合型搜索引擎的目的是把握全貌，而各种领域型搜索引擎则能够向用户提供深入、智能、更新及时的检索服务。每个搜索引擎都会就用户查询建议其他适合的搜索引擎。

由于不同的搜索引擎处理的信息，面向的用户群、发挥的作用都不用，它们分别适合采用不同的技术。结构分析技术适合应用于综合型搜索引擎；Metadata 演算、数据库、人工智能等技术一方面由于需要的人工劳动多、耗费资源大、要求原始数据规范性好，另一方面能够构造深入细致服务，适合于在领域型搜索引擎上应用。各种搜索引擎不再是各自为政，它们建立一个松散耦合的合作网络，以相对于整个合作网络最低的代价为用户的各种 IR 需求提供适当的服务。

事实上搜索引擎服务框架的核心思想不是一个全新的技术发明，当面临一个大型问题时，人们往往采用建立一个各司其职、相互合作的层次结构(Hierarchy)这种思路，如企业或国家的管理。搜索引擎服务框架也是我们在 Web 从完全的"无政府状态"向适当的有序化发展的一个设想和建议。

在我们的原型检索系统 SAInSE[1]中，我们尝试着对理想服务框架中的部分思想进行了检验。实验的效果是明显的。如果能够实现这样一个服务框架的话，那么对于 Web 上的查询，无论是查全，还是查准，无疑都是十分理想的解决方案。

参 考 文 献

1 马国臻. 基于结构分析的大规模 WWW 文本信息检索技术的研究:[2001中科院计算所博士论文]

2 Ricardo Baeza-Yates Berthier Ribeiro-Neto. Modern Information Retrieval. ACM Press, 1999

3 CLEVER. CLEVER Project, 1999

4 Collins, Collins J A, Schweitzer R. Applying metadata to the search interface: Constructing effective local and distributed searches of web-based scientific data. In: proc of the 5th WWW conf. 1994. On line at: http//www. cdc. noaa. gov/~jac/www. 95/paper. html

5 COOL links 1995 L. Jay Wantz, Micheal Miller. Towards user-centric navigation of the Web: COOL links using SPI. In: proc of the 6th WWW conf. 1995

6 COOL linds 1996 Michael Miller, L. Jay Wantz. COOL links: ride the wave. In: proc of the 7th WWW conf. 1996

7 Dean 1999 J. Dean, M. R. Henzinger. Finding related pages in the world wide web. In: 8th World Wide Web Conference, Toronto, May 1999

8 Silverstein 1998 Craig Silverstein, Monika Henzinger, Hannes Marias, Michael Moricz, Analysis of a very large AltaVista query log, DEC System Research Center (SRC) technical note, Oct. 1998

(上接第43页)

模型的支持，这种支持是通过一系列相关的类来实现的。对于 SAX 接口，在.NET 中也有相应的模拟实现。在 XML 数据源上，借助于 XML 提供的 DOM 接口或 SAX 接口，利用传统的数据挖掘算法即可进行 Web 数据挖掘，获取有用的知识，形成知识库。

我们已经利用 Microsoft VS. NET 的开发工具，实现了对若干农业网站上的农产品供求信息的 Web 数据挖掘。实践证明，本文提出的利用 XML 这种半结构化的数据模型辅助进行 Web 数据挖掘的方法是行之有效的。

结束语 Web 数据挖掘是一个新的很有前途的研究领域。由于 Web 中的数据是半结构化的，因此 Web 数据挖掘不同于传统的数据库中的结构化数据挖掘。如何针对 Web 上的数据为半结构化的这一特点，寻找一种半结构化的数据模型是 Web 数据挖掘必须解决的一个首要问题。当然，仅有这种半结构化的数据模型还不够，还需要有相应的半结构化模型抽取技术，即自动地从现有数据中抽取半结构化模型的技术。面向 Web 的数据挖掘必须以半结构化的数据模型和半结构化数据模型抽取技术为前提。XML 可看作一种半结构化的数据模型，它可以很方便地将 XML 的文档描述与关系数据库中的属性一一对应起来，实现精确的查询与模型抽取。

本文结合 IR 和 IE，提出的基于 XML 的 Web 数据挖掘方法已在实际系统中得到了成功应用。我们相信，此项研究一定会对 Web 数据挖掘起到积极的促进作用。同时，随着 XML 这一标准 Web 技术的不断发展，Web 数据挖掘的研究也必将取得越来越多的成就。

参 考 文 献

1 (加)Han J, Kamber M 著，范明，孟小峰 等译. 数据挖掘概念与技术(Data Mining: Concepts and Techniques). 机械工业出版社，2001

2 kosala R, Blockeel H. Web Mining Research: A Survey. ACM SIGKDD, July, 2000

3 Jackson J, Myllymaki J. Automatically extract information with HTML, XML, and Java. www-900. ibm. com , 2001

4 王超，张鹏. ASP. NET/XML 深入编程技术. 北京希望电子出版社，2002

5 徐振航，刘莉芹. XML 与面向 Web 的数据挖掘技术. www. ASP-Cool. com , 2001

6 刘振岩，王万森. 急切分类与懒散分类的研究. 小型微型计算机系统，2002待发