基于DNS的网页搜索引擎

华中科技大学
王亮
wl223@tom.com

报告大纲

- 一、搜索引擎遇到的主要问题
- 二、我们的解决方案
- 三、主要的问题和挑战
- 四、今后研究方向

一、搜索引擎遇到的主要问题

- 覆盖率。没有一个搜索引擎能够覆盖超过50%的互联网全部网页。
- 更新率。平均更新周期一个月。
- 检索结果的准确率问题。上万条检索结果意义并不是很大，且存在很多无关重复的检索结果。

问题根源

- 覆盖率和更新率问题：分布式的WWW和搜索引擎集中式结构之间的矛盾。网络等基础条件等方面的限制，WWW的动态特性，当前搜索引擎很难跟踪每处的变化。
- 准确率问题：HTML格式的不可读性以及人工智能发展滞后。

解决问题的方向和思路

- 覆盖率和更新率问题：采用地域上的分布式体系结构。
- 准确率问题：有待语义网和人工智能技术的发展

分布式搜索引擎研究

- 以harvest为代表的分布式搜索引擎研究。如东京大学的CSE，挪威科技大学的相关研究。均以服务器作为检索基本单位进行联合检索。
- 分布式搜索的两个难题：

1 要有合适的体系结构。目前P2P式体系结构难以保证检索质量和速度等基本要求，而过于集中又要遇到当前搜索系统的覆盖率、更新率等问题。

2 要有明确的实施需求和激励机制。作为一个分布式系统，其管理和建设必然是由不同的单位组织负责的，如果各个单位组织不能从系统的实施中受益，而仅仅是强调共享，技术再先进也只会是纸上谈兵。

体系结构选择问题

- 分散与集中式体系相结合可能是解决两种体系问题的关键。而DNS分层的分布式体系给了我们基本的启发。
- 如今几乎每个高校和大的机构都有自己的DNS服务器，并与高层服务器协调配合，这种分层的分布式体系使互联网上所有的站点都能得到有效的管理。
- 出于管理和效率等因素考虑DNS也经历了从集中式到分布式的转变。
- **结论：DNS这样的分布体系是否适合建立新型搜索引擎？既然DNS能够索引各个站点的名称，那么是否也能索引整个站点的所有网页呢？**

二、基于DNS的网页搜索引擎

1基本的结构体系

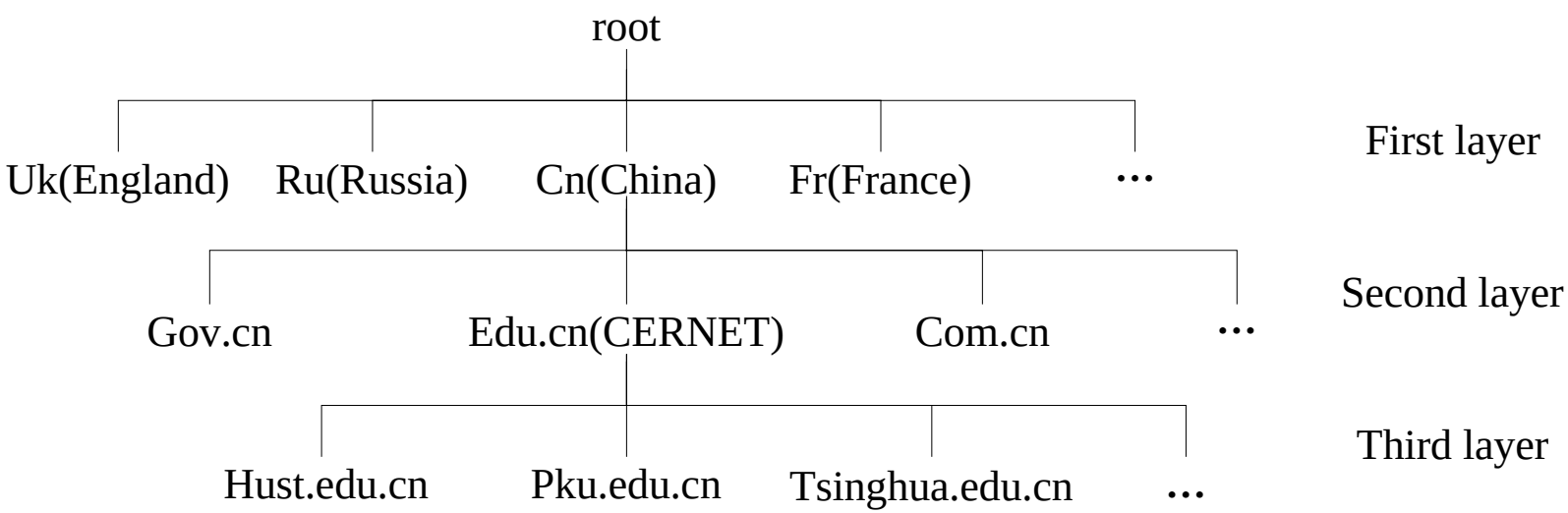


Figure 1 Architecture of system

2 基本实现思路

- 采用此基本框架，我们可以简单地在最底层下载网页数据，然后逐级传递到最上层的服务器上。
- 由于网页的下载更新工作都在不同的底层节点进行，而这些节点一般又都对应于某个局域网，因而这种分布采集、逐层递交的方式可以保证整个系统的数据每天更新，这样更新率问题就得到了很好的解决。
- 但是按照这种方法，顶层的服务器数据存储量可能依然很大，我们可能不得不采用分布计算等复杂技术来保障顶层服务器的数据存储和检索服务质量。
- 要建立一个可以“镜像”整个Internet数据的系统几乎是不可能的,必须采用其它方式来完成此任务。

2 相关的技术

按照基本体系结构划分，目前已有三种不同类型的信息检索系统：

1. **集中式检索系统。**这种系统拥有自己的数据采集装置，所有的数据都存储并索引在一个数据库系统中。如当前网页搜索引擎。
2. **元数据采集系统。**其采用从各个小的子数据库中采集元数据并整合到一个系统的方式构建信息检索系统。这类系统没有自己的数据采集模块，仅存储起索引功能的元数据，比较常用的如OAI系统。
3. **分布式检索系统。**分布式信息检索系统中各个子数据库系统分别提供符合统一标准的信息检索接口，执行信息检索时由总系统负责协调各个子数据源完成检索请求。著名的如Stanford数字图书馆计划中的InfoBus系。

- 信息检索系统基本结构的选择一般根据以下规则，即随着数据源规模扩大和数据类型的增多一般可以依次选择常规数据库型、元数据采集型、分布式检索型。

3 具体方案

- 按照DNS的基本结构，其按范围分为组织级、主干网级、国家级三级，数据量有少到多。
- 信息检索系统基本结构的选择规则:随着数据规模的增多可以选择集中式检索系统、元数据采集型、分布式检索系统。
- 我们将信息系统的基本结构选择规则应用到整个WWW上的信息管理，可以得到以下方法：**组织级——集中式检索系统，主干网级——元数据采集系统，国家级——分布式检索系统。**
- 在组织级别进行数据的采集和索引工作，然后向主干网级别的服务器提交元数据，而国家级的检索服务器则记录各个子数据源的检索接口描述数据。

第三层:集中式检索系统

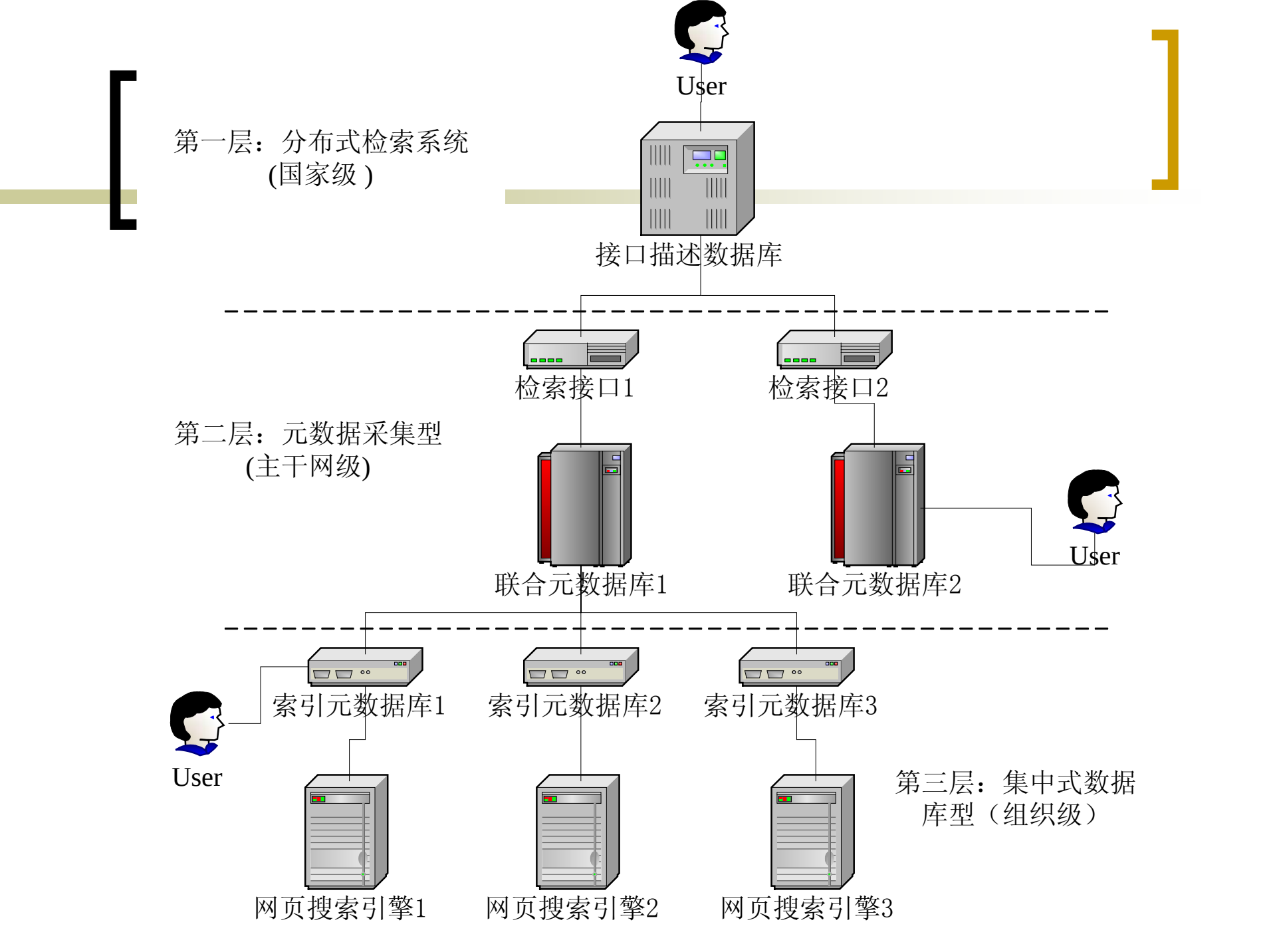
- 范围：组织单位内的本地局域搜索引擎。
- 工作原理：同一般网页搜索引擎。以服务器为单位逐个下载。
- 排序方式：同全文检索排序方式。
- 注意事项：当遇到指向其他服务器的链接时，也将此链接作为本站内容下载，但不再下载更深层次的链接，这样就保留了网页间的链接信息。

第二层：元数据采集系统

- 范围：主干网级别的搜索引擎，如CERNET上的网页搜索系统。
- 工作原理：元数据采集
- 排序方式：超链接分析。基于提交的同一网页数据的重复次数计算。

第一层：分布式检索系统

- 范围：国家级
- 工作原理：分布式检索
- 排序方式：元搜索的排序方式


User
第一层：分布式检索系统
(国家级)
接口描述数据库
检索接口1
检索接口2
第二层：元数据采集型
(主干网级)
联合元数据库1
联合元数据库2
User
索引元数据库1
索引元数据库2
索引元数据库3
User
第三层：集中式数据
库型（组织级）
网页搜索引擎1
网页搜索引擎2
网页搜索引擎3

4 系统特点和优势

	对应范围	系统结构	基本搜索技术	整体中的作用	存储内容
第三层	组织级	集中式	全文检索	下载器	原始数据
第二层	主干网级	元数据采集式	超链接分析	索引器	索引元数据
第一层	国家级	分布式	元搜索技术	检索接口	检索接口描述数据

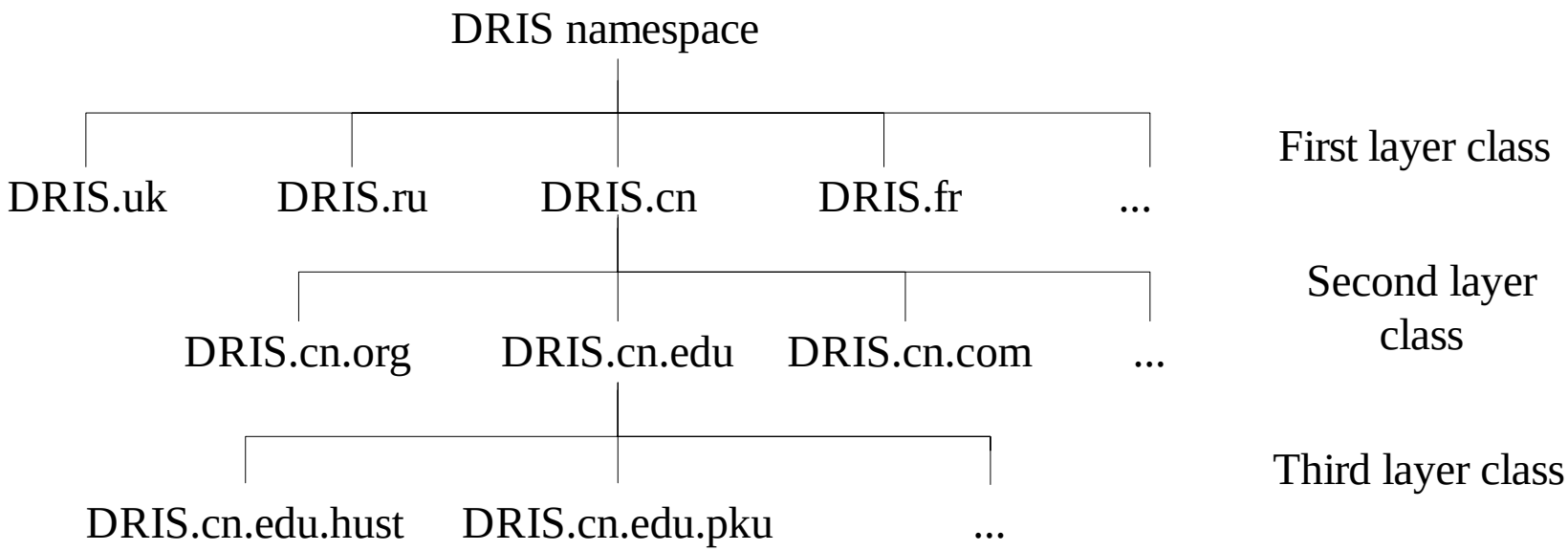
系统优势

- 覆盖率。由于本系统的下载器的工作是按域组织的，因此只要一个站点在域名系统中注册，其所有网页就可被新的系统索引，因而从理论上讲，基于DNS的检索系统可以覆盖所有互联网的网页。
- 更新率。新系统的网页下载和更新都在底层的各个服务器上进行，一般都对应于不同的局域网，其更新时间非常短，而在第二层，元数据上载过程也不用花费很多时间，而顶层由于没有实际的数据，因此不需要更新。所以整个系统的更新速度较现有系统有大幅度的提高。
- 准确率。由于新系统三层的每一个节点都是完整的搜索引擎，并可向外提供标准的检索服务接口，这就为很多个性化智能搜索系统提供了很好的数据源。在这样的个性化检索系统中，可以真正做到以用户为核心，这样的搜索结果显然会更精确。

5 系统应用协调和管理

- 由于新系统的每一个节点都是完整的搜索引擎，怎样使用户能够迅速找到需要的搜索服务是系统应用的关键。
- 利用用面向对象模型来描述此系统，使其成连为一个整体。
- 用Webservice/UDDI技术组织协调系统
- 我们为其选择一个了基本的命名空间“DRIS”

系统服务调用体系


DRIS namespace
DRIS.uk
DRIS.ru
DRIS.cn
DRIS.fr
...
DRIS.cn.org
DRIS.cn.edu
DRIS.cn.com
...
DRIS.cn.edu.hust
DRIS.cn.edu.pku
...
First layer class
Second layer class
Third layer class

基本规则

- 所有的节点都通过标准Webservice的形式提供检索服务。
- 所有的检索服务都按照“继承”的关系进行组织，低层的节点通过引用高层节点的Webservice的形式进行继承，高层节点通过一个专门的模块用来索引低层节点的检索接口。
- Web服务位置统一规则。Webservice通过URL链接来提供服务，每个DRIS服务器都通过链接“DRIS.域名”向外提供标准Webservice检索服务，而此服务器上Webservice的主类名为“DRIS.反顺序域名”。如华中科技大学的域名为 “hust.edu.cn”,则其DRIS服务器通过链接“DRIS.hust.edu.cn”向外提供校内各种资源的检索服务，而此服务的主类名为“DRIS.cn.edu.hust”。

三、主要的问题和挑战

标准化问题。作为一个公共搜索系统，标准协议的制定和实施关键。

- 要成为一个“事实上的标准”，就是要应充分考虑协议的可实施性和具体的推广工作。
- 要成为一个“权威的标准”，加强和IETF、W3C等标准化组织的联系合作。

实施问题。

- 采用此系统的激励机制何在？系统以“域”为基本单位的分层式结构为系统实施找到了基本的需求。在最底层建立了校园网/企业网内资源的网页搜索引擎，更高层系统可为地区/国家级基础信息平台建设提供方案。
- 由于其它域如.com下并不像.edu下那么组织有序，在系统建立时必须因地制宜，灵活地应用该系统的基本规则。例如有的国家地区可能所有的网页数据并不是很多，构建一种集中式的搜索引擎就可完全满足要求。

四、今后的研究方向

- 语义网研究。语义网的实施中的信息检索和网页设计之间的矛盾是否能找到一个折中的方式。
- 互联网信息基础体系研究。利用DNS的体系结构去整合其它类型的信息资源。“域内资源整合系统”正是为解决此问题提出的。

小结

主要工作

- 利用了DNS分层的分布式基本结构，确定了一种清晰的WWW信息管理系统基本结构，可以解决“覆盖率”和“更新率”问题。
- 在不同层次应用三种不同结构的检索系统，解决了“海量数据”的存储索引和管理问题。
- 利用Webservice/分布式UDDI体系将整个系统连成一个有机整体。为信息的智能化处理提供了基础性平台，为解决“准确率”问题提供了有效途径。

提供了一套较为完整的公共开放式的Web信息检索平台，并具有较好的可实施性。

END

THANKS !

Our site: http://dris.hust.edu.cn