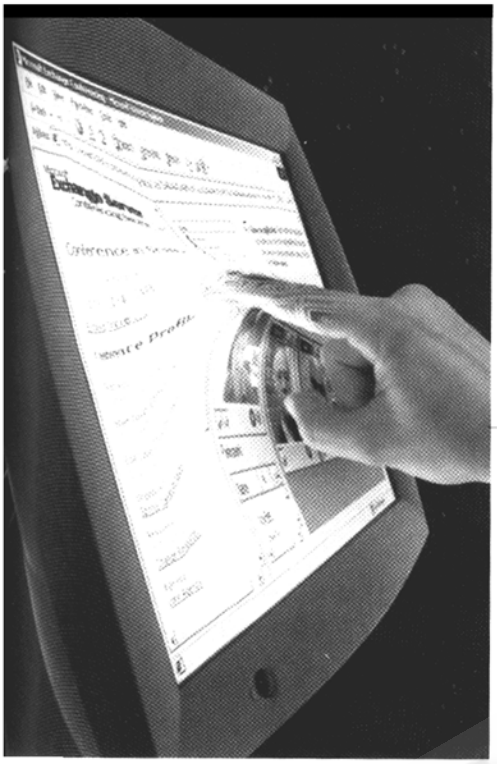

新一代面向XML网页搜索引擎的模型

卢 壮 (北京燕山石化研究院 102500)

摘要：随着XML技术的日渐成熟，利用XML发布网上信息已经逐渐成为一种趋势。面向XML网页的搜索引擎技术势必是下一个技术热点。然而，目前国内外尚无大型站点提供该种服务。基于该点事实，本文着重介绍未来面向XML网页搜索引擎的结构模型。

关键词：XML 搜索引擎 索引

基于内容的搜索引擎与基于文本的搜索引擎有很大的不同。前者不但要考虑关键词的匹配而且还要兼顾语义上是否一致，后者则简单得多，只要考虑字符的匹配与否就可以了。这就决定了基于内容的搜索引擎要比基于文本的搜索引擎复杂一些。

从目前的情况来看，这一类搜索引擎尚不多见，投入使用的还很少。但是它们一般都由三个部分构成：后台机器人模块，检索与搜寻模块和用户界面。机器人模块自动搜索XML文档，下载后交给检索与搜寻模块处理；检索与搜寻模块负责为这些文档建立索引，并且负责它们的检索和更新维护；用户界面主要是负责用户查询。其结构图如图1。

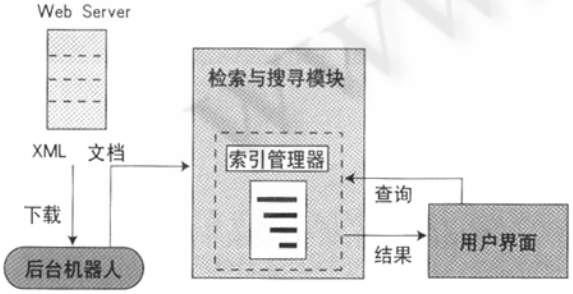

图 1

1 后台机器人模块

后台机器人模块主要负责下载列在尚未访问的URL队列中的文档，当文档被下载之后，该模块做两件事：一是在已经下载的文档中搜索链接，把这些链接放入到尚未访问的URL队列中；一是把该文档传给检索与搜寻模块，并把当前URL放入到已经访问的URL队列中。后台机器人模块并不负责回应用户的查询，它的任务只是遍历Web，并且把所有相关文档下载后传递给检索与搜寻模块。该部分一般来说可以分成检索、归类和过滤三个子模块。

1.1 检索子模块

在Internet上遍历是一件非常复杂的事情，单单进行单一搜索是不够的，因为网站都有一个更新的问题，如果没有重复遍历和及时修改的话，搜索引擎会变得陈旧。但是在这里不会过多涉及搜索算法的问题，我们假设搜索引擎只是进行没有修改的单一遍历，那么遍历的过程可以分为以下四步：

(1) 检索子模块由一个种子URL开始，假设它指向Web page，P；

(2) 检索子模块阅读P，把其中的链接提取出来放入尚未访问的URL队列中，设为L，然后把P放入到已经访问的URL队列中；

(3) 检索子模块把P传递给检索与搜寻模块；

(4) 检索子模块在尚未访问的URL队列中再取出一个URL，把它作为下一个种子URL，重复步骤1。

1.2 归类子模块

URL能够指定多种多样的种类的文档，如果不加限制的话，搜索引擎会把它们都加入到尚未访问的URL队列中。它们有些是有用的，需要为它们创建索引，而有些

是不必要创建索引的。这样，什么样的URL可以加入到尚未访问的URL队列中，便需要加以限制，这个正是归类子模块的功能。

1.3 过滤子模块

如果搜索引擎的所有者不加以节制的话，后台机器人模块的运行会带来大量的社会问题，包括对隐私和版权的侵犯。为了避免诸如侵权、服务器过载等问题的发生，搜索引擎必须遵守Standard for Robot Exclusion或者是其他的一些类似的防止搜索引擎侵犯网站协议。一般来讲，在每一个网站的根目录下应该有一个叫做robots.txt的文件。标准规定每一个搜索引擎都应该向站点提供诸如标示、搜索目的、与搜索引擎拥有者的联系方式等。每一次搜索引擎与站点进行连接时都应该提供此类信息。当搜索引擎工作的时候，其操作者应该监视其运行情况，以解决随时可能出现的问题。

2 检索与搜寻模块

检索与搜寻模块的主要任务就是为XML文档创建索引并且提供查询这些索引的方法。这部分并不负责与WEB连接和查询时所用的图形界面。后台机器人模块把XML文档传递给检索与搜寻模块后，它负责根据文档的内容结构为文档中的每一个词创建索引，这些索引由检索与搜寻模块中的索引管理子模块来进行管理，并由该子模块来负责索引的更新和查询。其结构图如图2：

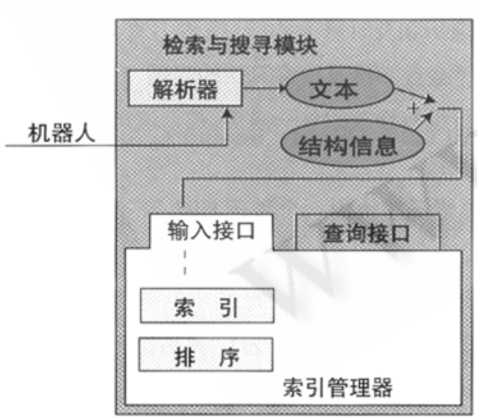

图 2

2.1 解析器

解吸器的任务是把XML文档区分成两个部分：一个是文档中的文本内容，一个是文档的结构信息。文档内容就是文档中的文本部分，而结构信息可以指明文档的背景信息和真实的含义。结构信息可以从封装文本内容的标记中得到。XML标记语言严格地维护标记的树状结构，它的语法与语义都是非常严格的，有清晰的起始和结束标记。所以通过文档中的标记来提取结构信息是比较容易的。

该部分实现的关键是区分文本内容与结构信息。相对XML来说，文本内容主要是指在标记之间的文本部分，如果是英文信息的话，可以通过词与词之间的空格轻松地提取每一个单词，进而传递给索引管理器；如果是中文信息的话，要对中文信息进行拆分，把其中的名词提取出来，然后再送给索引管理器。目前，关于中文信息的处理方法较多，较为常用的是分词词典法，即通过一个词典的支持对中文信息进行拆分。相对于文本内容来讲，结构信息的提取较为麻烦一些，该部分一般是通过扫描XML文档，分析标记的嵌套关系，从而得到某段文本内容的结构信息。

2.2 索引管理器

索引是通过索引管理器来维护的，它也负责在用户进行查询时进行信息定位。索引管理器有一个输入的接口，她负责判断解析器送过来的文本内容和与之相应的结构信息在索引列表中是否存在，如果文本内容已经存在而相应的结构信息不存在的话，那么可以把结构信息加到该文本索引的后面，如果连文本都不存在的话，那么就把文本与结构信息全都加进来。索引的结构如图3所示。

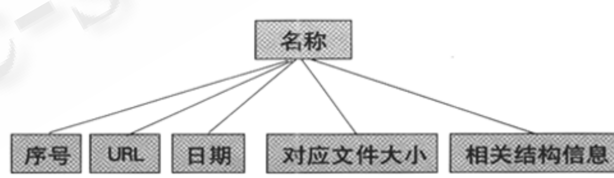

图 3

如果如上面所述，索引一定会非常冗长，大概是所对应文件的4倍。为了减少索引的大小，应该对索引进行转化。索引要包括文档的大小信息、索引日期，这些不宜压缩，而索引中的URL信息、文档的文本信息、文档的结构信息、日期信息等都是重复率比较高的，应该消除重复的数据。另外，在中英文中，冠词、代词、介词、连词等都没有必要建立索引。

3 用户界面

用户界面只是为索引管理器提供一个图形界面，简单、友好，大致与其他搜索引擎的用户界面相似。它并不分析文档、连接服务器或者创建索引。其结构图如图4：

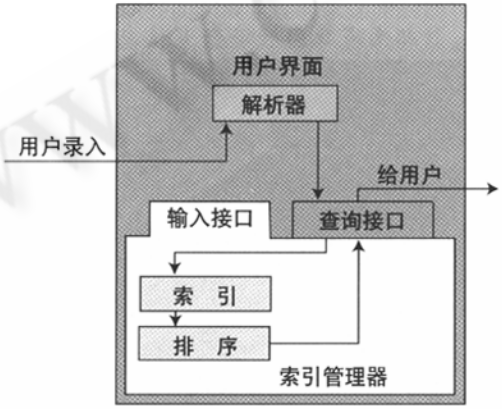

图 4

关于解析器。用户在查询信息时需要提供关键字和背景信息。对于用户的输入信息，解析器应该区分一下是否包含背景信息和逻辑表达式。然后它负责把用户输入翻译成索引管理器能够读懂的语法格式（包括逻辑表达式），用户的查询请求传递给索引管理器之后，索引管理器负责对索引进行查询，找到合适的URL，对它们进行排序后再传回到用户界面，形成查询结果输出。

4 结论

目前，基于XML的搜索引擎技术是未来Internet信息发掘技术发展的热点，也是下一代电子商务技术的主流之一。虽然面向XML主页的搜索引擎还没有进入实用的阶段，该模型也仅仅是处在一个起步阶段，正在不断修改，不过其应用前景势必十分广阔。■

参考文献

1 *What's the point of XML? Sun World Online: http://www.sun.com/sunworldonline/ swol-02-1998/swol-02-xml.html。*

2 *Tauber, J. A Comprehensive Introduction to XML. http://www.jtauber.com/xml。*

3 *XML 快速入门 电子工业出版社。*