基于最高响应比算法的WWW索引库更新方法

陈治平 林亚平
(湖南大学计算机与通信学院 长沙410082)

A WWW Indexed Database Refreshment Algorithm Based on the Highest Responding Ratio

CHEN Zhi-Ping LIN Ya-Ping
(The college of Computer and communication, Hunan University, Hunan 410082, China)

Abstract The problem is analyzed in refreshment of WWW search engine information database. With a view to the characteristic of WWW pages and effectiveness of information database, a new www indexed database refreshment algorithm based on the highest responding ratio is provided in this paper to relieve the heavy management of indexed information database.

Keywords Search engine , Highest responding ratio, Indexed information database

1 引言

WWW搜索引擎(Search Engine)利用网络蜘蛛收集WWW上的相关文档信息,通过分析、处理后,将相应的文档信息加入本地信息库,在用户给定其所关心的查询条件后,利用WWW搜索引擎所提供的检索查询系统从信息库中检索出符合用户要求的信息列表,并计算每条信息与用户要求的查询条件的相关程度,按照倒排序的方式返回给用户,使用户能够快速定位到他所关心的信息[1]。由于搜索引擎提供了这样一种工具,使得用户可以在众多的网页信息中能够快速定位到他所要求的信息上去,因此,搜索引擎应用是Web应用中最广泛的应用之一。

由于Web网页的动态变化的特性,使得以前被搜索引擎所搜索并进入索引库中的网页信息可能由于信息的更新或网址的变动而造成这些信息不再存在,这些错误信息或无用的信息提交给查询用户时形成垃圾信息,从而导致搜索引擎的整体性能下降。因此,搜索引擎需要不断地对其索引信息库能够进行更新维护,以保证索引信息的有效性。

搜索引擎通常采用单一周期信息更新策略,每隔一段固定的时间间隔对索引信息库进行一次信息更新维护过程。两次信息更新之间的时间间隔通常以星期或月为计量单位,如WISE系统的更新时间为两星期[2],CORA系统的更新时间为两个月[3]。

对于一个长期运行的搜索引擎来说,有效性验证是系统更新维护过程中一项十分重要而又费时的工作。一般大型的搜索引擎其索引信息库中都包含有几千万个Web地址网页信息,如果在每次更新过程中,对于每个网页都进行有效性验证,这个工作量将是十分巨大的。假设对于单个网页的一次有效性验证所需的时间是0.001秒,则要完成几千万的Web网页的验证工作也需要几天的时间。而事实上,有许多文档内容是很少改变的,没有必要不断地验证其有效性。另一方面,由于一个搜索引擎只有一个固定不变的信息更新周期,而且这个周期通常都比较长,一般都是以星期或月为单位,因此对于一些变化比较频繁的文档,其索引信息的时效性就难以得到充分的保证。

基于上述原因,本文提出了一种基于最高响应比的信息更新算法。该算法利用索引信息的检索情况,以及网页的更新频度计算每类网页的更新周期,按照更新周期的情况进行信息的有效性检查,从而不仅减轻了系统信息维护的工作量,而且保证了信息的有效性。

2 搜索引擎工作原理及其信息更新过程

搜索引擎的整体结构如图1所示。

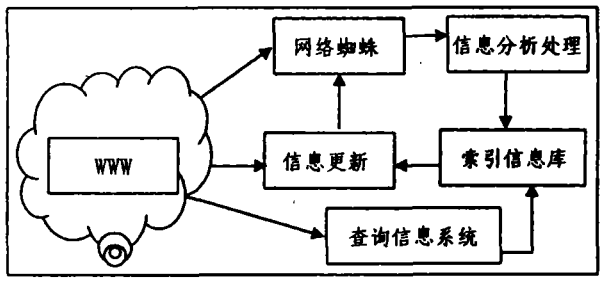

图1 搜索引擎整体结构

网络蜘蛛自动从网络上获取网页信息,并从获取的网页文本信息中抽取出该网页所连接的其他网页地址信息,作为蜘蛛下次将要获取的网页信息的URL地址,通过这样的方式,可以访问互联网上的所有信息。

网络蜘蛛将获取的网页信息进行信息处理,按照预先规定的算法进行分析,测试该网页内容与搜索引擎搜索目标的相似程度,将符合要求的信息及其相关信息送入索引信息库中以便查询系统使用。

查询信息系统通过其提供的查询接口以某种预定的形式提交给用户,用户输入相应的查询条件后,将信息提交给查询系统,查询系统根据给定的查询条件检索索引信息库,并计算每条信息与用户要求的查询条件的相关程度,按照倒排序的方式返回给用户。倒排序的结果使得最满足用户要求的信息出现在最前面,从而达到快速定位的目标。

由于Web网页的动态变化的特性,使得以前被搜索引擎所搜索并进入索引库中的网页信息可能由于信息的更新或网址的变动而造成这些信息不再存在,这些错误信息或无用的信息提交给查询用户时会导致查询失败,或其他错误结果,使搜索引擎的整体性能下降。因此,搜索引擎需要不断地对其索

陈治平 博士,讲师,主要研究方向:智能搜索引擎和数据库技术。林亚平 博士导师,主要研究方向:计算机网络、机器学习。

引信息库进行更新维护，以保证索引信息的有效性。这部分的工作由信息更新系统完成信息的更新过程。

索引信息库的信息更新过程通常包括两个部分：验证信息库中已有连接的有效性；更新从上次维护结束以来内容发生变化的文档和新出现的文档信息。

有效性验证根据文档的最新修改时间，搜索引擎通过发送 HTTP HEAD 请求来获取文档的最新修改时间，通过比较修改时间的不同来判断内容是否发生了变化。如果服务器的响应表明文档已不能访问，则从索引信息库中删除对应的记录；如果时间不一致，则说明文档的内容可能发生了至少一次的变化情况，则把文档地址加入到一个特定的目标列表中。这个目标列表将用来启动网络蜘蛛进行第二步（新一轮）文档信息收集过程。

搜索引擎通常采用单一周期信息更新策略，每隔一段固定的时间间隔对索引信息库进行一次信息更新维护，对于一些变化比较频繁的文档，其索引信息的时效性就难以得到充分的保证。

基于这种认识，本文提出了一种基于最高响应比的信息更新算法。该算法利用索引信息的访问情况，以及网页的更新频度计算网页的更新周期，按照更新周期的情况进行信息的有效性检查，从而不仅减轻了系统信息维护的工作量，而且保证了信息的有效性。

3 基于最高响应比算法的更新方法

定义1（系统检测周期 P） 搜索引擎连续两次的索引信息有效性验证之间的时间间隔。

定义2（文档检索度 $A(d_i)$） 自上次系统进行有效性验证以来对于某个文档索引信息 d_i 的检索次数。

规则1 文档检索度大，该文档被用户访问的概率也比较大，对于这样的文档索引信息，如果在文档内容发生变化时不能得到及时的更新，会降低系统的整体性能，因此必须确保这样的文档的有效性，使得文档一旦发生变化能够得到及时的更新。

定义3（文档稳定度 $C(d_i)$） 搜索引擎对文档 d_i 的最后一次更新时间到本次准备进行有效性验证时的时间间隔与系统检测周期 P 的比率；

$$C(d_i)=(Now - Time_{Last\ Modified})/P \quad (1)$$

规则2 文档稳定度越大，相应文档的变动的可能性越小。

定义4（文档变动率 $M(d_i)$） 系统检测周期 P 与搜索引擎对文档 d_i 的前两次连续的索引信息变动之间的时间间隔的比率；

$$M(d_i)=\begin{cases}0 & \text{若该文档索引没有变动，或者}\\ P/\text{前两次变动时间间隔} & \text{若该文档索引发生过变动}\end{cases} \quad (2)$$

规则3 文档变动率越大，相应文档的前两次连续的索引信息变动之间的时间间隔越短，则该文档变动的可能性也越大。

定义5（文档检测响应比 $\mathscr{R}(d_i)$） 搜索引擎对该文档 d_i 的有效性检查的响应程度

$$\mathscr{R}=(\lg(A(d_i)+1)+M(d_i))/C(d_i) \quad (3)$$

由公式(3)可知：文档检索度 $A(d_i)$ 越大，相应的文档检测响应比 $\mathscr{R}(d_i)$ 也越大，因此，对于检索次数比较高，访问的概率比较大的文档信息可以通过文档检索度来保证索引信息的有效性；文档变动率 $M(d_i)$ 越大，相应的文档检测响应比 $\mathscr{R}(d_i)$ 也越大，这对于信息更新比较频繁的索引信息文档也可以使得有效性能够得到保证；另一方面，文档稳定度 $C(d_i)$ 越大，文档变动的可能性也越小，而对于这些文档信息而言，由于它们的文档检测响应比 $\mathscr{R}(d_i)$ 比较小，获得系统更新的能力比较弱，因此可以达到降低系统开销的要求。

通过研究 Web 网页的特性，我们发现大量的网页在信息确定后信息修改的频度非常低，基本上不会进行任何修改，结合这个特点，设定系统有效性检查的文档检测响应比阈值或设定系统有效性检测的百分比方法，给定文档检测响应比阈值方法只对检测响应比高于阈值的文档进行更新，而对于低阈值的索引信息我们不进行更新；设定百分比方法从索引信息库挑选出属于该百分比范围的高响应比索引信息进行更新。这样，我们在保证系统整体性能不下降的前提下，只对极少部分的文档索引信息进行更新，因此，系统的开销也就大大降低了。

4 基于最高响应比方法的算法描述

根据上面的分析，我们给出如下的基于最高响应比方法的算法：

```
Begin
  初始化响应比队列列表;
  For(I:=0;I<文档信息数量;I++)
    {计算第 I 个文档索引信息的响应比;
     将计算出来的响应比按由小到大的顺序插入响应比队列中;}
  按照预定的方式挑选出待检查的索引信息;
  For(I:=1,I<待进行有效性检查的文档索引数;I++)
    {验证第 I 个待检查的索引信息;
     IF 信息不存在 THEN 从索引库中删除该索引信息;
        ELSE 更新信息,并更新修改时间;
    }
```

首先利用公式(3)计算出各网页的响应比，使用二分法查找算法将响应比的大小按从大到小进行排序。

根据预定的方式挑选出待进行有效性检查的索引信息。预定的方式可以为给定阈值或某一特定百分比的值进行。具体数值可以根据应用系统的需要进行相应调整。由于索引信息数目是时刻在变化的，给定一个具体数值不符合应用系统的要求，因此我们主要采用百分比的方法进行更新。

更新待检查的索引信息，并对真正需要更新的索引信息在进行更新索引的同时修改信息库中的相应记录内容，作为下次有效性检测的依据。

从算法中，我们可以得出算法的复杂度为 $O(N\log_2 N+N\times$给定的百分比$)$。百分比是个小于1的数目，在 N 比较大的情况下我们可以知道 $\log_2 N \gg 1$，因此，算法的复杂度可以简化为 $O(N\log_2 N)$。

结束语 WWW 搜索引擎信息库在进行信息更新的过程中一般都没有考虑到对信息更新的优化问题，从而使得信息更新维护开销比较大；同时过长的时间更新周期又会降低系统的性能，为了在保证系统整体性能不受影响的前提下降低系统开销，特提出了一种基于最高响应比算法的 WWW 索引信息库更新方法，既结合了大量 WWW 文档的稳定的特性，同时考虑到变化的索引信息更新的时效性，从而减轻了索引信息库维护的工作负担。

参 考 文 献

1 王继成，萧嵘，孙正兴，张福炎. Web 信息检索研究进展. 计算机研究与发展[J]，2001，38(2)：187～193

2 Yuwono B，Lun D. WISE：A World Wide Web Resource Database System [J]. IEEE Transactions on Knowledge and Data Engineering [J]，1996，18(4)：548～554

3 McCallum A K，Nigam K，Rennie J，Seymore K. Automating the Construction of Internet Portals with Machine Learning. Information Retrieval Journal，2000，3：127～163