领域相关的 Web 网站抓取方法*)

李 刚 周立柱 郭 奇 林 玲
(清华大学计算机科学与技术系 北京 100084)

摘 要 本文提出了一种抓取领域相关的 Web 站点的方法,可以在较小的代价下准确地收集用户所关心领域内的网站。这种方法主要改进了传统的聚焦爬虫(Focused Crawler)技术,首先利用 Meta-Search 技术来改进传统 Crawler 的通过链接分析来抓取网页的方法,而后利用启发式搜索大大降低了搜索代价,通过引入一种评价领域相关性的打分方法,达到了较好的准确率。本文详细地描述了上述算法并通过详细的实验验证了算法的效率和效果。

关键词 Meta-Search ,聚焦爬虫(Focused Crawler),启发式搜索

Website Crawling for Specific Topics

LI Gang ZHOU Li-Zhu GUO Qi LIN Ling
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)

Abstract In this paper, we propose a new approach to discover the Websites for special topic in WWW with high precision and low cost. This approach improves traditional Focused Crawler techniques, different from the common Web crawler which accesses the Web graph composed by HTML pages and hyperlinks, our crawler uses Meta-Search to get the URLs of relevant page, then uses heuristic search method to reduce the search cost, and uses topic relevant rules to increase the precision. The experimental results show the presented approach is both effective and efficient .

Keywords Meta-Search ,Focused crawler, Heuristic search

1 简介

随着网络的迅速发展,万维网已经成为大量信息的载体,如何有效地提取并利用这些信息已经成为了一个巨大的挑战。通用的搜索引擎如 Google, Yahoo, 百度等,已经成为了人们应用互联网信息的主要工具和入口。但不同领域、不同背景的用户往往具有不同的检索目的和需求,因而如何满足不同领域用户的特定需求,成为了现在搜索引擎领域的一个重要的问题。

同类型网站是指具有同样一种性质,在某一个特定领域内,对用户提供服务的网站,例如提供网上购书服务的网站。如何通过尽可能小的代价来寻找这一种类的网站,传统的方法是采用改进 Crawler 技术的 Focused Crawler 方法[1]。传统 Crawler 技术是从一个或若干初始网页的 URL 开始,在抓取、分析网页的过程中,不断地从当前页面上抽取新的 URL 放入队列,直到满足系统的停止条件。Focused Crawler 的工作流程类似传统 Crawler,但它根据一定的搜索策略从队列中选择下一步要抓取的 URL[2]。图 1 是传统的 Crawler 和 Focused Crawler 的工作示意图。

但 Focused Crawler 在对此类网站的抓取中存在着一个非常重要的问题,就是 Focused Crawler 本质上仍然属于传统的 Crawler 技术,也就是通过链接分析来进行整个互联网网页的抓取,于是网页之间的链接关系就对整个的抓取过程产生至关重要的影响。

但在互联网中,同类型网站存在着商业竞争关系,所以这些网站彼此之间基本上不存在直接的链接关系,即使有可能通过其他网站彼此相连,这些中转网站也可能属于类型不相关的网站[3],所以会对 Focused Crawler 的抓取过程产生了巨大的不良影响。通过对 500,000 个网页的调查,Bergmark[4] 发现大多数属于相关类型的网页被不相关的网页所割裂,它们的距离在 1 和 12 之间,平均距离是 5。Kumar[5,6] 通过考察 100,000 个网页也得到了类似的结论。Masashi Toyoda[7] 对这种网络聚合体(Web Community)的问题有比较全面的论述。图 2 就是这种问题的示意图。

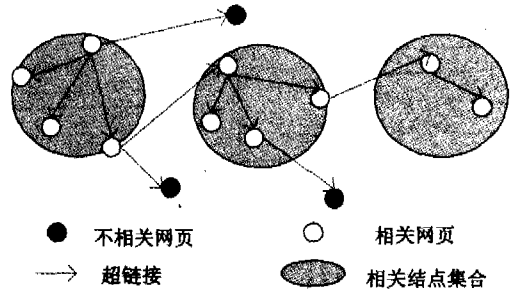

图 1 传统 Focused Crawler 工作示意图

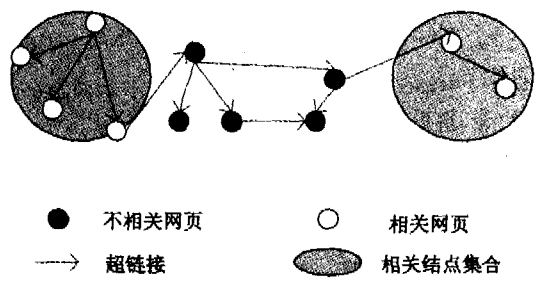

图 2 传统 Focused Crawler 对于同类型网站抓取问题示意图

现在对于同类型网站需求主要是通过人工进行收集、整

*)基金项目:国家自然科学基金重大国际合作项目:超对等语义搜索引擎,2004-2006,项目编号:60520130299。

理的方式进行解决，即编制相应的导航型页面，例如 Alex 排名 63 位的站点 www . hao123. com，就是一个这样的导航型页面。它将互联网存在的主要站点进行了人工的分类整理，类似传统的电话黄页方式。由于人工收集整理是一种效率十分低下的方法，而 Focused Crawler 又有一定的局限性，因此在这篇论文中我们提出了一种代价非常小的收集同类型网站的方法。

主要思路是：首先通过具有该领域特点的查询利用搜索引擎得到大量的具有一定类型相关性的网站，接着利用一种启发式的搜索方法依次检验 Meta-Search 的搜索结果，最终得到所需要的同类型网站。本文后面就是对这种方法的阐述，第 2 部分介绍了有关 Focused Crawler 和领域搜索的相关工作，第 3 部分是整个算法的详细说明，第 4 部分通过实验来证明算法的有效性，最后是全文的总结和一些未来工作的展望。

2 相关工作

Focused Crawler 概念最初提出的目的就是在互联网中抓取特定主题的网页，所以也叫 Topic Crawler。Soumen Chakrabarti 的论文[1]是具有代表性的聚焦爬虫的早期研究之一，目前大多数的聚焦抓取都采用了类似的工作流程，其系统结构如图 3 所示。根据一个主题目录和用户指定的初始点来描述抓取目标，并在用户浏览过程中，将用户标注的感兴趣网页放入相应的主题目录，修改主题样本。系统的两个主要部分是网页分类器和网页选择器(distiller)。网页分类器负责学习抓取目标的特点，计算网页的关联度，并过滤网页。选择器负责计算网页的重要程度，发现中心型网页，并由此动态决定网页的访问顺序。

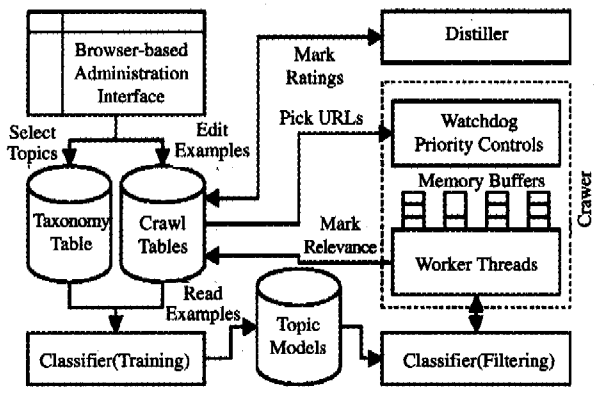

图 3 说明网页分类器和网页选择器如何结合的 Focused crawler 的方框图

而后 Focused Crawler 的发展主要集中在网页的扩展方法与链接和页面领域相关度分析方法两个方面。

对于网页的扩展方法，主要有深度优先、广度优先和最佳优先三种[2]。深度优先在很多情况下会导致爬虫的陷入(trapped)问题，目前常见的是广度优先和最佳优先方法[8]。最佳优先搜索策略按照一定的网页分析算法，预测候选 URL 与目标网页的相似度，或与主题的相关性，并选取评价最好的一个或几个 URL 进行抓取。它只访问经过网页分析算法预测为“有用”的网页。存在的一个问题是在爬虫抓取路径上的很多相关网页可能被忽略，因为最佳优先策略是一种局部最优搜索算法。因此需要将最佳优先结合具体的应用进行改进，以跳出局部最优点[9]。研究表明，这样的闭环调整可以将无关网页数量降低 30%～90%。

对于网页和链接的分析方法，除了改进经典的链接分析算法 PageRank 和 Hits 算法以外，主要是基于文本和链接的内容进行领域相关性的考察。聚焦抓取常以三种方法表示：(1)预给初始种子样本(如种子 URL，目标网页样本等)；(2)预定网页分类结构(如 yahoo!)和网页训练集生成的分类器；(3)用户显式标注的或从日志推理得到的“有用”样本。

三种方法都只是对抓取行为的“主题性”或所关心的“领域”给出了模糊的定义。文[9]采用了预定义的本体信息，领域本体由不同的概念、实体及其之间的关系，以及与之对应的词汇入构项(lexical entry)组成。网页中的关键词在通过与领域本体对应的词典作规范化转换之后，进行计数和加权，算出与所选领域的相关度。北京大学的 Commix 系统[10]在计算相关度上也应用了类似的方法。

3 算法介绍

整个系统主要是分为两部分：Meta- Search 部分与启发式网站检验部分。具体的流程如下：

1)将指定的关键字发送给网络上的搜索引擎，进行查询，得到一系列拥有关键字的网页列表，将所得到的所有查询结果作为下一阶段的输入。

2)对网页列表中的网页对应的网站利用启发式的搜索算法依次进行检查，判断是否为所需要的相关类型站点，最终得到一系列相关类型网站的列表。

具体的流程图如图 4 所示。

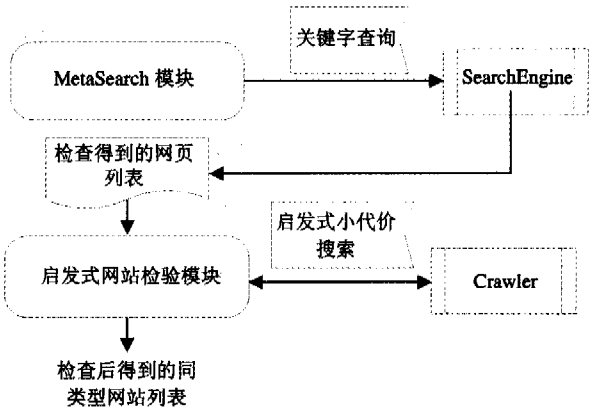

图 4 算法流程示意图

3.1 Meta-Search 部分

这部分是利用搜索引擎，通过搜索包含所需数据信息的关键字，得到大量的网页列表。其主要目的是得到一系列具有一定相关性的网页，这个过程可以看作一个粗略的数据采集过程。

发送给搜索引擎的关键字要与所需采集的网站有一定的相关性。由于现在搜索引擎功能的强大，而且后边会仔细地检验每一个网页所对应的网站，因此这个相关性可以十分模糊，其目的是得到具有一定相关性的大量的结果，是一个粗略的筛选过程，以作为数据集提供给后边的网站检查模块。

当然，不同的关键字会对搜索结果有很大的影响，后边有不同关键字对搜索结果影响的实验。从我们一般的搜索经验以及实验中可以看出，最好的方式是采用多种关键字进行搜索，然后对所有搜索的结果采用合并的方式产生最全的网站

列表。由于后边检查的算法代价非常小，因此在 Meta-Search 部分中，更强调的是“全”这个方面，而对“好”这方面的要求相对减弱。

3.2 启发式网站检验部分

利用启发式搜索方式，对 Meta-Search 部分搜索得到的所有网站一一进行检查，判断是否为所需要的同类型的网站。采用启发式搜索的目的是尽量减少验证代价（即验证过程中实际访问的网页数量）。

3.2.1 网页类型的相关性评价方法

首先定义网页 P 与所要收集类型关系的函数 $F(P, R)$，R 是一个评价规则集合，每个规则包含两个部分：

规则描述：规则描述是基于正则表达式的与、或、非逻辑表达式。

权重：满足规则后得到的相应分数，是归一化以后的结果。

例如在进行网上书店领域相关网站的抓取时，用来描述价格属性和属性值的表达式为，“（市场价|售价|原价|定价|现卖价）(:)? [0-9.]+($|¥|元)?”，其权重为 0.263。

其表达式的含义为网页中应该包含这样的字符串，首先出现“市场价”、“售价”、“原价”、“定价”、“现卖价”中的一个字符串，然后出现一次或零次冒号，然后出现一个或多个数字组成的数字串，最后出现单位符零次或一次。

例如以下字符串均满足规则：“市场价：21 元”，“定价 32.20”，“现卖价：12 $”。

通过此函数可以对某一个网页 P 进行打分，得到此网页与搜集网站类型的相关性 $F(P,R)$ $(0\leqslant F(P,R)\leqslant 1)$。在我们所进行的搜集网上卖书网站的实验中，共定义了“价格信息”、“网上购买信息”、“出版信息”、“作者信息”、“书名信息”、“版次信息”、“简介信息”以及“ISBN 信息”共八条规则。

3.2.2 AbstractURL 的概念

AbstractURL 目的是按照 URL 的特性对网页进行分组，如果同网站的两个网页具用同样的 AbstractURL，则假设具有同样收集相关性分数。

对于某一个网页 p_1 来说，AbstractURL(p_1)的定义如下：

一个 URL 一般的组成为：http://[Host][Path][File][Query]? [Ref] ?

例如：p_1 的 URL 为 = http://www.china-pub.com/computers/result.asp? typeid= C24-04-01

则我们定义的 AbstractURL 为：

AbstractURL=[Path][File][Query 中的属性名]

对于 p_1 来说，它的 AbstractURL 就是 /computers/result.asp? typeid。

AbstractURL 的意义在于同网站内的网页可以按照 AbstratURL 进行分类，而具有相同 AbstractURL 的网页可以认定是属于同一类型的数据，有很大的可能性具有类似的收集相关性分数($F(P, R)$)，特别是在“同类型网站收集”中主要关注的数据性网站。

另外如果定义关系 R 为“AbstractURL 相同”，则此关系 R 对相同网站网页的分类是一种等价分类，可以很容易地知道，由 AbstractURL 定义的相同关系满足等价分类要求的自反性，对称性和传递性。所以从理论上来讲，我们如果知道了某一个网站 W 的全部网页集合 $P = \{p_1, p_2, p_3 \cdots\}$，则可以得到通过 AbstractURL 相等关系来划分的等价类集合 $A = \{A_1, A_2, A_3, \cdots\}$。

3.2.3 启发式的最小代价网站验证算法

检查网站的过程实质上是一个启发式搜索的过程，即利用 AbstractURL 的值对此网站中的网页分类后进行预估分数，对预估分数较高的 AbstractURL 类中的未扩展网页进行优先扩展。

定义启发函数值：对于具有同样 AbstractURL 的页面集合 A_k，其中已经扩展并计算分数的页面为 E_k，则启发函数 $H(A_k)=1/|E_k| * \sum(F(p_i,R))(i\in E_k)$。这个启发函数意义是：此等价类集合中已扩展网页的相关分数平均值。

具体的算法如下：

Input：初始未扩展的网页 p_0。

Output：p_0 所属的网站 W 是否为相关网站的判断，TRUE 表示相关网站，FALSE 表示不是相关网站。

A：等价类的集合，A_i 表示其中的第 i 个等价类，A_i 中的网页拥有相同的 AbstractURL。A_i 由已扩展的页面集合 E_i 和未扩展的网页的集合 Q_i 组成。即 $A_i=E_i+Q_i$。

$[p_i]$：网页 p_i 对应的 AbstractURL 等价类，A 是$[p_i]$的集合。

PageThreshold：分数阈值，为初始参数，当一个网页 p 的相关性分数值大于 PageThreshold 时（即 $F(p, R) >$ PageThreshold），被认为是相关的网页。

NumRelPages：相关网站阈值，即网站包含的相关网页大于 NumRelPages 时，此网站被认为是好网站。

NumExpPages：扩展网页数量的阈值，即网站已经扩展的网页数(Ncheck)大于 NumExpPages 时，此网站被认为是坏网站。

```
Begin
A←{{p0}}
i←0
numGoodPage ← 0
Do
  i←i+1
  B={Ai| Ai 中含有未扩展的网页}
   if(B≠φ)then
   begin
    find Ak:Ak∈B 且 对于Ai∈B,H(Ak)≥ H(Ai)
    find pj:pj∈Qk
    扩展 pj，即利用 Crawler 抓取 pj 的页面
     if(F(pi,R)≥ PageThreshold )
     begin
      numGoodPage ← numGoodPage +1
      if ( numGoodPage ≥ NumRelPages )
          return TURE
     end-if
     H(Ak)← (H(Ak) * |Ek| + F(pi,R))/ ( |Ek| +1)
     Ek← Ek ∪ {pj}
     Qk← Qk-{pj}
    N←页面 pj 中属于网站 W 但没有被扩展过的新链接集合
     for pi∈N do
      if ([pi]∈A)then
        [pi]←[pi]∪{pi}
      else
        A←A∪{{pi}}
      end-if
    end
    else
      return FALSE
    end-if
until (i≥ NumExpPages )
return FALSE
End.
```

4 实验

实验主要是考察各种参数对整个算法的影响。算法的结果主要由三个参数影响，即 PageThreshold，NumRelPages 和 NumExpPages。

实验中所寻找的相关类型网站是提供网上购书服务的网站，所使用的搜索引擎为 www.baidu.com。

4.1 各种参数对算法准确率的影响

准确率的定义是：

准确率＝实际上相关的网站/评判为相关的网站＊100%

其中分数阈值 PageThreshold 对准确率的影响最大，不同的分数阈值对网页的判断结果有很重要的影响，分数阈值越高就越准确，但由于一些相关网站的页面信息不是很全，因此分数阈值 PageThreshold 越高，发现的相关网站就越少，图 5 反映了分数阈值 PageThreshold 对准确率的影响。

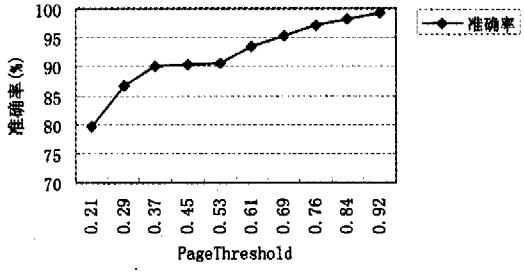

图 5　PageThreshold 对准确率的影响图

图 6 是分数阈值 PageThreshold 对发现的相关网站数目的影响图。

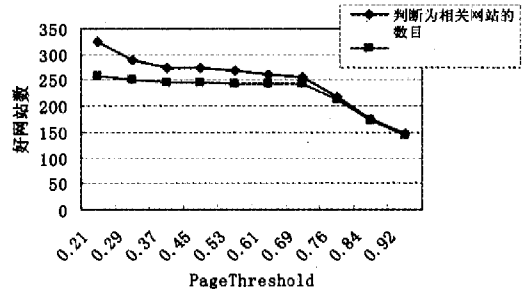

图 6　PageThreshold 相关网站数目影响图

从以上的数据可以看出，分数阈值对准确率影响是十分巨大的，随着分数阈值的增大准确率不断增大，但大于 0.7 以后由于条件逐渐苛刻，相关网站数目有大幅度减少，因此分数阈值取在 0.7 左右可以得到一个比较好的结果。

另外两个参数 NumRelPages 和 NumExpPages 对准确率的影响不大，这两个参数主要影响的是搜索代价。

4.2　各种参数对抓取代价的影响

整个算法的开销主要集中在网络抓取，这是因为相对于从 Web 中抓取网页，算法所需要的本地的磁盘 IO 和 CPU 计算可以忽略不计，所以性能由抓取的网页数量决定，这里用平均网页数量来衡量，即：

平均抓取网页数量＝总共抓取的网页数量/检验的网站数量

相关网页数量阈值 NumRelPages 对平均抓取网页数量的影响(查询词为“网上书店”)如图 7 所示。

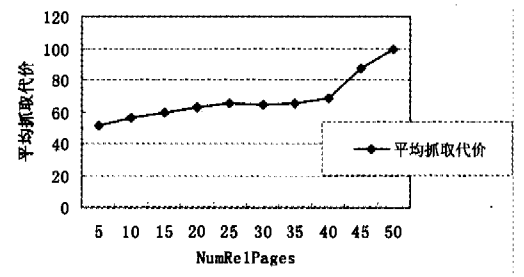

图 7　NumRelPages 对代价的影响图

扩展网页数量阈值 NumExpPages 对平均抓取网页数量的影响(查询词为“网上书店”)如图 8 所示。

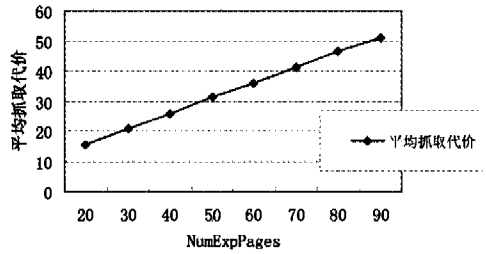

图 8　NumExpPages 对代价的影响图

从实验中可以看出，NumExpPages 对平均抓取代价有非常大的影响，基本上是线性增加的，而 NumRelPages 在不是很大的时候对代价的影响不大。

从上边的实验数据可以看出，代价与准确率的关系不是十分明显，这是因为准确率主要由相关性评价算法所决定，而与网页的抓取数量关系不是很明显。

4.3　不同领域查询词对算法的影响

搜索相关率在一定程度上可以反映关键字的优劣，其定义是：

搜索相关率＝相关网站数量/搜索到的网站数量＊100%

实验考察了网上购买图书这个领域的情况(表 1)。

表 1

查询词	查询词说明	网站数量	相关网站数量	搜索结果的相关率(%)
网上书店	比较普通的领域用语	613	258	42.088
作者 定价	属性＋数据特征	616	208	33.766
程序设计 定价	实例查询	575	135	23.478
购物车 书	领域用语＋概念	592	253	42.736
购物车 作者	领域用语＋属性	497	192	38.632
作者 定价 购物车 书	多种关键字的组合	614	368	59.935

实验同样考察了提供网上订购手机这个领域的情况(表 2)。

表 2

查询词	查询词说明	搜索到的网站数量	相关网站数量	搜索结果的相关率(%)
网上购物 手机	比较普通的领域用语	575	263	45.739
外观 市场价	属性＋数据特征	603	148	24.544
诺基亚 定价	实例查询	475	89	18.737
购物车 手机	领域用语＋概念	595	195	32.737
购物车 像素	领域用语＋属性	423	142	33.570
定价 像素 购物车 手机	多种关键字的组合	598	255	42.642

(下转第 148 页)

ICDE Conference, 2001
2 Abiteboul S. Querying Semistructured data. In: Proceedings of the International Conference on Database Theory, 1997
3 Suciu D. Semi-structured data and XML. In: Proceedings of International Conference on Foundations of Data Organization, 1998
4 聂培尧,李战怀,等. 基于 OEM 的 XML 半结构数据的模式描述方法. 计算机工程与设计, 2003,24(1)
5 Abiteboul S, Quass D, McHugh J, et al. The Lorel Query Language for Semistructured Data. International Journal on Digital Libraries, 1997,11(1)
6 Clark J. XML Path Language. http://www.w3.org/TR/xpath
7 Deutsch A, Fernandez M, Suciu D. Storing Semistructed Data with STORED. In: Proceedings of the ACM SIGMOD International Conference on the Management of Data, 1999
8 Nestorov S, Ullman J, Wiener J, et al. Representative Objects: Concise Representations of Semistructured Hierarchical Data. In: IEEE International Conference on Data Engineering, 1997
9 Abiteboul S, Buneman P, Suciu D. Data on the Web: From Relations to Semistructured Data and Xml. Morgan Kaufmann, San Francisco, 1999
10 Mchugh J, Widom J, Abiteboul S, et al. Indexing Semistructured Data: [Technical Report]. Stanford University Computer Science Department, 1998
11 Goldman R, Widom J. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. In: Proc. of the 23rd VLDB Conference, 1997
12 Milo T, Suciu D. Index Structures for Path Expressions. In: Proc. of the 7th International Conference on Database Theory, 1999
13 高军,唐世渭,等. 半结构化数据查询重写. 计算机研究与发展, 2002, 39(2)
14 Tseng V, Lin W. A New Method for Indexing XML Documents. In: Proc. of the 12th Workshop on Object-Oriented Technology and Applications, 2001
15 Cooper F, Sample N, Franklin M, et al. A Fast Index for Semistructured Data. In VLDB, September 2001
16 Grahne G, Mendelzon A O. Tableau Techniques for Querying Information Sources Through Global Schemas. The MIT Press, 1999
17 Pottinger R, Levy A. A Scalable Algorithm for Answering Queries Using Views. In: Proc. of VLDB Journal, 2001
18 Papakonstantinou Y, Vassalos A. Query Rewriting Using Semistructured Views. In: Proceedings of the ACM SIGMOD International Conference on the Management of Data, 1999
19 Calvanese D, De Giacomo G, Lenzerini M, et al. Rewriting of Regular Expressions and Regular Path Queries. In: Proc. of PODS'99, 1999
20 Calvanese D, De Giacomo G, Lenzerini M, et al. Containment of Conjunctive Regular Path Queries with Inverse. In: Proc. of KR 2000, 2000
21 Grahne G, Thomo A. Algebraic Rewritings for Optimizing Regular Path Queries. Theoretical Computer Science, 2003,11(6)
22 Fernandez M, Suciu D. Optimizing Regular Path Expressions Using Graph Schemas. In: IEEE International Conference on Data Engineering, 1998
23 Park S, Kim H J. SigDAQ: An Enhanced XML Query Optimization Technique, Journal of systems and software, 2002,61(2)
24 Florescu D, Kossmann D. Storing and Querying XML Data Using an RDBMS. IEEE Data Engineering Bulletin, 1999
25 Calvanese D, De Giacomo G, Lenzerini M, et al. View-based Query Processing and Constraint Satisfaction. In: Proc. of LICS 2000, 2000

(上接第 140 页)

从实验中可以看出如下几种选取查询词的规则:

1)比较普通的关键字已经可以带来很好的搜索结果:这是由此算法的特殊性所决定的。因为算法中的 Meta-Search 部分主要起一个初步筛选的过程,得到较多相关结果为此部分的主要目的,所以普通的关键字更能得到多的结果。

2)多种关键字的组合可以得到较好的结果:这是因为可以更加详细的描述所要查询的相关网站,所以可以得到更加多的相关网站数量。

3)整个算法对查询词的依赖性不是特别强烈。只要选择比较具有普遍领域知识的查询词,都可以得到比较好的结果。

4)在实际应用中,由于每个网站只检查一次,结果中重复的网站不会对算法造成影响,因此可以采用多种关键字的组合来进行多次的搜索,网站的检查列表采用每组关键字检索结果的并集,会得到比较好的结果。

4.4 实验总结

从整个实验可以看出,算法的结果还是比较好的。搜索"网上书店"关键字得到的 627 个网站中发现 233 个相关的提供网上购书服务的网站,准确率达到 95.28%,而平均抓取代价只有 31 个网页,总共访问了 19716 个网页。这个抓取代价是非常小的,因为一般情况下一个普通的购书网站的网页量就可以达到几万甚至几十万的数量级。用普通的 Crawler 在普通的 PC 机上一小时即可抓取所需要的网页,完成整个实验。

表 3

站点名	Alex 排名	相关网站数目	准确率(%)
www.hao123.com	64	22	91.667
www.265.com	91	65	84.415
www.mscncn.com	55065	86	68.254
我的算法	NA	222	95.27

而相对应的导航型站点,由于是人工整理,其结果的数量以及准确率都不高,以下是对一些导航型站点的数据。表 3 是一些比较数据(以下实验的实验日期是 2006 年 3 月 1 日,Alex 的排名和网站的情况都以当日的情况为准)。

从中可以看出,导航型站点提供的相关网站由于是人工整理的原因,其数量和准确率都不是很好。而本文的算法无论是在准确率还是在数量上都达到了较好的效果。

总结 本文介绍了一种抓取领域相关型网站的小耗费算法,主要阐述了此算法的具体过程以及利用此算法进行的实验。从实验中可以看出,算法在非常小的代价下取得了很好的结果,发展了传统 Focused Crawler 的算法,无论是相关网站发现的数量或是质量上均大大超过了现在的人工收集方法。

未来进一步的工作是算法如何在不同领域中自动生成领域相关的网页评价规则和查询词,而不是由人为设计。这将是一项长期而有意义的工作。

参 考 文 献

1 Chakrabarti S, van den Berg M, Dom B. Focused crawling: a new approach to topic-specific Webresource discovery. In: Proc. of the 8th International World Wide Web Conference, Toronto, Canada, 1990
2 周立柱,林玲. 聚焦爬虫技术研究综述. 计算机应用,2005(9)
3 Qin Jialun, Zhou Yilu, Chau M. Building Domain-Specific Web Collections for Scientific Digital Libraries: A Meta-Search Enhanced Focused Crawling Method. In: JCDL'04, 2004
4 Bergmark D, Lagoze C, Sbityakov A. Focused Crawls, Tunneling, and Digital Libraries. In: Proc. of the 6th European Conference on Digital Libraries, Rome, Italy, 2002
5 Kumar R, Raghavn P, Rajagopalan S, et al. Extracting Large-Scale Knowledge Bases from the Web. In: Proc. of the 25th International Conference on Very Large Data Bases Conference, Edinburgh, Scotland, UK, 1999
6 Kumar R, Raghavan P, Rajagopalan S, et al. Trawlling the Web for Emerging Cyber-Communities. In: Proc. of 8th International World Wide Web Conference, Toronto, Canada. Machine Learning Techniques, Proc. AAAI-99 Spring Symposium on Intelligent Agents in Cyberspace, 1999
7 Toyoda M, Kitsuregawa M. Creating a Web Community Chart for Navigating Related Communities, In: Proc. 8th WWW Conference, 1999
8 Cho J, Garciamolina H, Page L. Efficient crawling through URL ordering [A]. In: Proceedings of the Seventh International Conference onWorldWideWeb [C], April 1998
9 Ehr I, Maedche A. Ontology2focused crawling of Web docu ments [A]. In: Proceedings of the 2003 ACM symposium on App lied omputing [C], March 2003
10 韩近强,赵静,杨冬青,唐世渭. 基于领域知识的网页筛选系统. 见:第 19 界全国数据库学术会议论文集,计算机科学,2002(8)