第 27 卷　第 4 期　　华侨大学学报(自然科学版)　　Vol. 27　No. 4
2006 年 10 月　　Journal of Huaqiao University (Natural Science)　　Oct. 2006

文章编号　1000-5013(2006)04-0429-04

面向侨务信息主题的搜索引擎系统

吴清江　吴　政　刘琳琅

(华侨大学信息科学与工程学院, 福建 泉州 362021)

摘要　介绍面向侨务信息主题搜索引擎的工作原理和体系结构, 根据侨务信息的特征对网页进行侨务信息的识别. 针对主题式搜索, 提出一种优化的, 基于历史反馈(BHF)的搜索策略, 并对该搜索策略进行实验测试. 结果表明, 以该方法设计的面向侨务信息主题的搜索引擎系统, 具有较高的搜索速度与识别精度.

关键词　搜索引擎, 历史反馈, 侨务信息, 主题式搜索, 搜索策略

中图分类号　TP 393.09　　文献标识码　A

随着 Internet 应用的迅速发展, 侨务信息的载体也从原来的单一的报刊杂志逐渐转移到了 Internet 上. 侨务工作者希望能够迅速、方便地从 Internet 上收集大量的、分散零碎的侨务信息. 设计一个能自动从 Internet 上收集侨务信息的系统, 是侨务信息处理中亟待解决的工作问题. 本文给出了一个基于历史反馈(BHF)的搜索策略, 并在此基础上设计了面向侨务信息主题的搜索引擎系统. 该系统能够在 Internet 上搜索和自动收集相关的侨务信息.

1　系统工作原理

面向侨务信息主题的搜索引擎工作原理, 如图 1 所示. 搜索引擎在搜索引擎策略控制器的控制下, 将相关网页下载下来, 提交给信息预处理模块进行处理; 预处理模块主要完成编码转换, 文本提取等工作, 随即进入侨务信息网页识别模块. 该模块采用匹配处理方式判断网页对于侨务信息的相关程度, 如果相关程度大于某个临界值, 则认为是侨务信息而进行存储. 另外, 搜索引擎模块的存储队列也通过存储模块进行存储. 最后, 侨务信息查询与提取模块从存储数据库中提取相关的信息资源呈现给用户. 搜索引擎的体系结构由搜索引擎模块、信息预处理模块、网页的侨务信息识别模块、搜索引擎策略控制器模块、储存模块和侨务信息查询与提取等组成. 从图 1 可以看出, 侨务信息源来自于 Internet, 当用户访问面向侨务信息主题的搜索引擎时, 向搜索引擎提出检索要求, 搜索引擎则从 Internet 中检索出符合要求的信息存入数据库, 同时将这信息展示给用户以供选择.

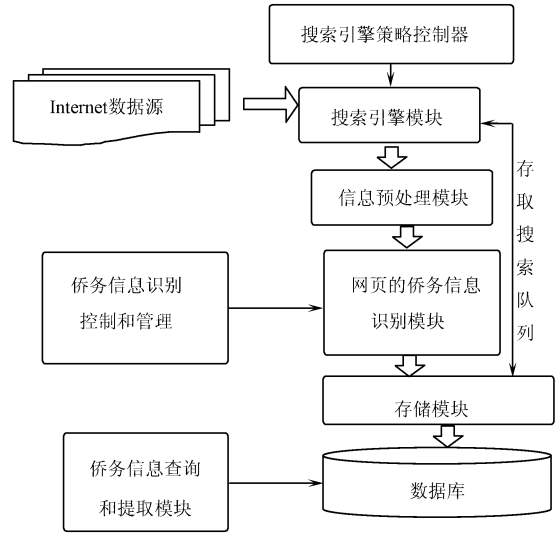

图 1　工作原理图

通常情况下, HTML 格式的文件采用的是“ISO-8859-1”的编码方式, 而搜索引擎被设计成运行在中文系统环境中, 所以必须对下载下来的 HTML 数据实现从“ISO-8859-1”的编码方式到“GB2312”的编码方式的转换[1]. 由于很多程序语言都提供了这种编码转换的功能, 本文就不再赘述. HTML[2] 数据

收稿日期　2006-05-22

作者简介　吴清江(1949-), 男, 副教授, 主要从事网络数据库及图像处理的研究. E-mail: wuqingjiang@msn.com

基金项目　国家计划委员会重点科研基金资助项目(ZX2000)

中的文本是我们处理的对象,因此必须剔除HTML格式中的标签而保留文本.在实际处理文本时,需根据标签的意义,把标签分为两类:一类是起分隔作用的标签;另一类是不起分隔作用的标签.如果两个文本之间出现不起分隔作用的标签,则应认为它们是连续的.

2 侨务信息的识别

侨务信息网页的识别,采用如下2种方式.(1) 关键词匹配.这种方式是最为普遍和简单的识别方式,而且识别速度非常快.但它在对综合性信息进行处理时,往往因为没有上下文的语义理解,造成严重的"答非所问"现象.本文研究的是面向侨务信息主题的搜索引擎,虽然也没有上下文的语义理解,但歧义发生的可能性要小得多,一方面因为侨务信息具有其特有的关键词,如华裔、侨办、侨联、……;另一方面,我们还可以采取其他避免歧义的措施.(2) 关键模式匹配.对某些侨务信息,仅用关键词匹配是不够的,因为部分侨务信息的特征是以一定的模式存在的.使用模式匹配方式对侨务信息进行识别,扩大了侨务信息覆盖范围,增加了侨务信息量.

侨务信息的关键词和关键模式,主要由侨务专家根据其经验收集来的.我们首先求出侨务信息的关键词或关键模式权重W.设有m个样本侨务信息文档,分别为$P_1, P_2, \cdots, P_m$;有n个侨务信息关键词或关键模式,分别为$K_1, K_2, \cdots, K_n$.通过对样本文档进行关键词或关键模式的匹配,统计出数据为

$$\begin{matrix} C_{11} & \cdots & C_{1n} \\ \vdots & \vdots & \vdots \\ \vdots & 0 & \vdots \\ C_{m1} & \cdots & C_{mn} \end{matrix}.$$

其中,C_{ij}为关键词或关键模式K_j在文档P_i中出现的次数.设W_j为关键词或关键模式K_j的权值.定义$\sum_{j=1}^{n} W_j C_{1j} \geqslant 1$, $\sum_{j=1}^{n} W_j C_{2j} \geqslant 1$, …, $\sum_{j=1}^{n} W_j C_{mj} \geqslant 1$.从上述不等式可以看出,$W$不是唯一的,只要取其中的一组就行.在求$W$时,要注意$W$一定是大于0的,且任一不等式的右边不能大于1太多.

设有一个待识别的文档P,通过关键词或关键模式匹配处理,得出$C_1, C_2, \cdots, C_n$.其中,C_j为关键词或关键模式K_j在文档P中出现的次数.设$t = \sum_{j=1}^{n} W_j C_j$.如果$t \leqslant 1$,代表该文档与侨务信息无关;$t \geqslant 1$,代表该文档与侨务信息有关.我们定义侨务信息网页相关度r为

$$r = \frac{t}{t+1}, \quad t \geqslant 0,$$

r代表了该网页与侨务信息的相关程度.当$t = 1$时,$r = 0.5$.所以,如果$r \leqslant 0.5$,代表该文档与侨务信息无关;如果$r \geqslant 0.5$,代表该文档与侨务信息有关.

3 搜索策略

在本文中,我们采用两种搜索策略结合的方式[3].一种是从一个或多个已知的站点出发,通过网页中的链接,不断地扩大搜索范围的搜索策略;另一种是BHF搜索策略.第1种搜索策略实现起来比较简单,通过不断地下载网页,提取链接即可实现.下面,我们对第2种策略进行分析[4].

定义**1** 侨务信息网页链接相关度总和l_r,表示该网页上包含的所有网页链接所指的网页的侨务信息网页相关度r_j的总和.即$l_r = \sum_j r_j$.

定义**2** 侨务信息新网页链接相关度总和l_m,表示搜索引擎在一次搜索该网页时,包含的所有与侨务信息相关的,且以前未曾搜索到的最新网页链接所指的网页的侨务信息网页相关度r_i的总和.即$l_{rn} = \sum_i r_i$.

定义**3** 侨务信息旧网页链接相关度总和l_{ro},表示搜索引擎在一次搜索该网页时,包含的所有与侨务信息相关的以前已搜索到的网页链接所指的网页的侨务信息网页相关度r_j的总和.即$l_{ro} = \sum_j r_j$

另一个重要的概念是,侨务信息网页链接更新率r_{ef},即

$$r_{ef}(n) = \frac{l_{rn}(n)}{l_{ro}(n)}, \qquad n > 0$$

上式中, n 为第 n 次搜索, $r_{ef}(n)$ 为第 n 次搜索该网页的更新率, $l_{rn}(n)$ 和 $l_{ro}(n)$ 分别为第 n 次搜索该网页时, 该网页的侨务信息新网页链接相关度总和和侨务信息旧网页链接相关度总和.

当搜索引擎刚开始搜索时, 由于没有历史文档, 只能进行盲目搜索. 这时, 虽然搜索的范围很广, 但获取侨务信息的效率却很低下, 并且搜索周期也较长. 因此, 搜索路径只能是顺着被搜索的页面上的链接搜索下去, 而这些链接所指的网页是否与侨务信息有关, 却不得而知. 搜索引擎试探性地将其下载下来后, 再进行侨务信息的识别. 在 Internet 上, 侨务信息相对于 Internet 上的海量信息来说是很少的. 如果搜索引擎只是试探性地盲目搜索侨务信息的话, 获取侨务信息的效率将极低, 要想及时获取最新的侨务信息, 必须优先对一些权威性高, 更新较快的网页进行搜索. 当搜索引擎经过了两个以上的搜索周期后, 获取的信息已经足够多了, 这时就可以采用 BHF 搜索策略. 通过上面的公式可以看出, $r_{ef}(n)$ 为第 n 次搜索该网页的侨务信息新网页链接相关度总和与侨务信息旧网页链接相关度总和的比值. 如果网页的更新比较快的话, 所包含的最新侨务信息也会较多, $l_{rn}(n)$ 也就会相对较大, 而 $l_{ro}(n)$ 就会相对较小. 那么, 在第 n 次搜索时, 它的 $r_{ef}(n)$ 也会较大.

侨务信息网页链接更新率, 充分反映了一个网页包含了侨务信息相关的网页链接的更新程度. 如果让搜索引擎优先对侨务信息网页链接更新率大的网页进行搜索, 就可以及时地从这些网页中获得较新的侨务信息. 在第 n 次搜索时计算出的 $r_{ef}(n)$, 实际上被用于第 $n+1$ 次搜索的判断. 因为在第 $n+1$ 次搜索完成之前无法得到 $r_{ef}(n+1)$, 也就是说, 第 n 次搜索的结果被用于指导第 $n+1$ 次搜索. 搜索引擎按 $r_{ef}(n)$ 从大到小的顺序进行搜索, 并且搜索时, 仅对它的子链接进行搜索, 也就是说只进行 1 层搜索, 而不进行更深层次的搜索.

BHF 搜索策略支持更快地获得最新信息, 但它却受历史文档的限制. 因为搜索引擎是由历史文档来指导进行搜索的, 已有的历史文档限制了搜索引擎的搜索范围. 因此, 在这种情况下, 我们不排除进行盲目搜索. 一种有效的做法是让一部分线程进行盲目搜索[5], 而另一部分线程则进行 BHF 搜索. 盲目搜索使搜索的范围更大, 一些新建的网站和曾经因为某种原因(如临时关机等) 而为未被搜索到的网页不断加入进来, 这样能及时地获得最新信息. 本文所研究的搜索引擎具备了两者的优点, 互为补充, 使搜索效率更高.

4　实验与测试结果

4.1　侨务信息识准率和识全率

我们收集了 1 000 个网页, 其中, 因为用于测试, 有 200 个是经过人工挑选的侨务信息网页; 其他 800 个非侨务信息的网页主要来自搜狐、新浪等门户网站, 并且经过人工校对为非侨务信息网页. 用本系统对其进行了侨务信息的识别, 结果如下: 识别出是侨务信息的网页为 201 个, 其中真正属于侨务信息的网页为 193 个. 由此可知, 识准率为 96.02%, 识全率为 96.50%.

通过分析, 我们发现不属于侨务信息却判断为侨务信息的网页中, 含有一些歧义性的词, 如"余华文选"、"华文雅思培训"、"北京歌华文化发展"、"华侨中学"等. 对于这类问题的解决办法是, 构造排除关键词字典, 如果排除关键词字典表中的词足够丰富, 识准率会进一步提高. 对于属于侨务信息, 却没有被判断为侨务信息的网页, 这主要是因为关键词的权值存在一定的误差, 解决这种问题的方法是使用更大数目的样本侨务信息网页. 另外, 我们还可以刻意选一些有这种问题的网页加入到样本网页中. 采取以上措施后, 识全率也可以进一步提高.

4.2　侨务信息获取效率

因为本文采用的是 BHF 搜索策略, 它是基于前一次搜索的结果进行的. 如果让其在 Internet 上运行, 可能需要很长时间才能计算出侨务信息获取效率. 为了测试侨务信息获取效率, 我们自行模拟 Internet 环境. 实验方法和结果描述如下: 构造出 1 000 个网页, 这些网页按照一定方式相互链接, 其中有 50 个是与侨务信息相关的. 在这些网页中选出几个起始网页, 然后进行搜索测试. 此时采用的是盲目的搜索方式. 搜索引擎搜索完这 1 000 个网页, 耗时大约 317.34 s, 其中搜索到 50 个与侨务信息相关的

网页,耗时276.35 s.接下来,将50个与侨务信息相关的网页中的25个更换为新的,与侨务信息相关的网页,让搜索引擎同样采用盲目的搜索方式,搜索完1 000个网页大约为326.18 s,则搜索到这25个与侨务信息相关的新网页的时间为201.56 s.然后,我们再次更换25个侨务信息相关的网页,让搜索引擎采用BHF搜索策略进行搜索,搜索到这25个与侨务信息相关的新网页的时间为8.93 s.结果表明,采用盲目搜索方式搜索25个与侨务信息相关的新网页,其搜索效率p_1为0.124个·s^{-1};而采用BHF搜索方式搜索25个与侨务信息相关的新网页,其搜索效率p_2为2.80个·s^{-1}.从以上数据可以得出,侨务信息获取效率提高了p_2/p_1= 22.58倍.

5 结束语

本文介绍了面向侨务信息主题的搜索引擎的工作原理和体系结构,阐述了侨务信息网页的识别方法,提出了BHF的搜索策略,并对以该策略技术设计的搜索引擎系统进行了测试.通过与其他搜索方法的比较、分析,可以看出,BHF的搜索方法明显优于其他方法.该搜索引擎可满足侨务信息主题资源检索的要求.

参 考 文 献

1 Coffman E G, Liu Jr Z. Weber R R. Optimal robot scheduling for Web search engines[J]. Journal of Scheduling, 1998,(6):14~22

2 Mark A C, Overmeer J. My personal search engine[J]. Computer Networks, 1999,31(21):2 271~2 279

3 Zamir O, Etzioni O. Grouper: A dynamic clustering interface to Web search results[J]. Computer Networks, 1999, 31:1 361~1 374

4 杨 杰,徐炜民.搜索引擎技术的运用与研究[J].计算机工程,2002,(1):265~272

5 洪光宗,王 皓.搜索引擎Robot技术实现的原理分析[J].现代图书情报技术,2002,(1):99~101

Study Search Engine for Information about Overseas Chinese Affairs

Wu Qingjiang Wu Zheng Liu Linlang

(College of Information Science and Engineering, Huaqiao University, 362021, Quanzhou, China)

Abstract In this paper, we first introduced the work principle and the framework structure of topic-special search engine for the information about overseas Chinese affairs. We found a way to identify the information of the overseas Chinese affairs on the suitable web page based on the characters of the information about the overseas Chinese affairs. By this topic-special searching, the paper brought forward an optimization search strategy which is based on historical feedback (BHF). At the end, we took an experiment on this search strategy. The results show that, this search engine system has a high search speed and identify precision.

Keywords search engine, historical feedback, information about overseas Chinese affairs, topic-special searching, search strategy