情报检索

# Internet 中文信息搜索引擎 GoYoYo 的研究

陈朝晖

(中国科学院文献情报中心 北京 100080)

【摘要】 介绍了 Internet 网络中文信息搜索器 GoYoYo 的基本情况,并深入研究了该搜索器的使用方法与技巧,同时客观地评价了它的优劣得失,指出需进一步完善之处。

【关键词】 Internet 搜索引擎 情报检索 中文信息

## Studies on Chinese Search Engine GoYoYo over Internet

Chen Zhaohui

(*The Documentation and Information Center of the Chinese Academy of Sciences*)

【Abstract】 On the authors' experiences, this article introduces the basic backgrounds of the Chinese Search Engine GoYoYo over Internet, researches the method and skills of its usage, and objectively estimates the GoYoYo, points out what should be optimized.

## 1 简 介

目前 Internet 上有很多优秀的信息搜索引擎,如 Yahoo、Infoseek、AltaVista 等,可以说是一个巨大的电子图书馆的检索室,但是它们只限于英文资料,对于英文不太好的用户来说,查阅起来有一定难度。现在专门用于查找中文信息的搜索引擎出现了,它就是 GoYoYo(中文名为"悠游搜索器",URL:http://www.goyoyo.com.hk/),由美国优联克国际有限公司(Unilinx International USA)于 1997 年 5 月中旬在香港推出。GoYoYo 是一个高度智能的中文搜索引擎,它已和全球 24 万个中文互联网网页相连,它的智能机器系统以两天为周期不停地搜索全球无数个互联网网页,查找新网页和网页中的最新资料,并自动进行识别和分类,自动转换繁、简体字。用户既可以采用关键词,也可以使用分类途径进行检索,找到自己所需要的信息,更可藉着相关页的索引,进入其它相连的网址,相当方便。因此,自 5 月 15 日正式推出以来,每天访问该网址的约 1 万多人次。

GoYoYo 的分类途径包括保健、财经、地区、电脑、教育、科技、社会、时事、文史、艺术、娱乐及政治等十二个主类及其各小类。

其收录范围为 24 万个中文页。为了节省存储空间,本站只收录中文网页。那些网页中仅含有个别的汉字,或者其内容与我们收录的内容相差很大的均不在该站搜罗之列。

## 2 系统配置、功能特点

GoYoYo 采用优联克公司生产的 Emplink 900 Internet/Intranet 中文网络服务器,其服务器系统配置包括:Web Server(页服务器)、FTP Server(文件传输服务器)、E-mail Server(电子邮件服务器)、DNS Server(域名服务器)、Internet Gateway(国际互联网网关)、Security

收稿日期:1997-08-25

图 1　主题分类表

Controls (系统安全控制)、PPP 点到点协议拨入设备、硬盘空间租赁、自选域名等。

使用 Emplink-900, 设置 Intrnet/Interanet 服务器变得容易而迅速。在不到 30 分钟的时间里, 就可以预置好 Internet/Intranet 服务器。

其优点:

1) 容易预置: 备有快速入门教程;

2) 容易连接: 支持所有路由器。为 PPP 拨号用户的设备提供了一个 HTML 格式的接口;

3) 容易管理: 支持本地和远程管理。可为任何基于 PC、Macintosh 或 UNIX 工作站的网络浏览器提供一个 HTML 格式的接口。

该服务器的所有软件功能都已预置完毕, 只需键入使用参数即可使用。

其功能特色主要有以下几个方面:

1) 带有 EmpView-900 型 Internet/Intranet 服务器操作系统, 奔腾平台及可连接的设备;

2) 功能集成化;

3) 灵活的用户自选设置, 支持域名缓冲分解及 E-mail 路由选择;

4) 可利用任何网络浏览器容易地进行网络管理;

5) 采用了 UNIX 的平台来保证数据的安全;

6) 无需任何 UNIX 操作指令, 所有的工作由鼠标点按的方式来完成;

7) 整套的保安系统包括: 用户口令、网络管理员口令及目录进入口令;

8) 灵活的 Internet 连接——工业标准化的 IP 路由器 PPP 拨号用户支持;

9) 工业标准化的以太网接口, 包括内部局域网的连接;

10) 记录文件可用于分析网上文件来往及 E-mail 的活动情况;

11) 自动文件备份或预置时间的自动文件备份系统;

12) 灵活的 E-mail 名字自选方式;

13) 提供多种语言的用户界面等。

## 3　GoYoYo 搜索引擎的特点

GoYoYo 搜索引擎由以下几部分组成: 1) 中文的预处理, 包括分词与词性标注; 2) 万维网网页的收集; 3) 网页关键词的抽取; 4) 数据库的检索; 5) 用户的万维网界面。

GoYoYo 具有与其它的搜索器不同的特点: 1) 它不是一个固定网络检索系统, 不需要完全依赖网页持有人把网址增加在其资料库中, 它除了有增加网页功能外, 主要是主动出去搜寻, 自动将新的网页添加进来; 2) 它不须等待有经验的系统操作员把网页分类, 而是依靠其超智能机器系统作精细的词汇区分; 3) 具有关联性网络资料索引功能, 使它突破传统搜索器的局限, 自动在网页文章内找出关键字或词, 在按用户要求去进行搜索时, 能把合适的文章和有关联性的文章一并找出。举例来说: 读者要找有关"音乐会"的文章, Goyoyo 会把与"演奏会"、"演唱会"、"小提琴演奏会"、"钢琴演奏会"、"莫扎特"、"贝多芬"等有关的文章也找出来, 满足读者的要求; 4) 具有精密的中文词自动切分与词性标注功能, 搜索中文网页准确快捷; 5) 一个复合算法(Hybrid approach)系统, 使智能机器系统具有学习功能, 并且大大提高了系统的性能; 6) 中文信息的自动分类和信息过滤; 7) 自动收集处理英文、中文国标码及大五码的网页, 用户使用支持其中一个中文码字的浏览器, 便可查出全部网页, 并且阅读其全部内容; 8) 用户无须使用空格把词分开, 也无须使用布尔逻辑算符和其它符号, 可完全按照书写习惯输入检索请求, 它主要利用词与词之间的语义联系进行主题检索, 而非进行简单的字串匹配, 方便易

用。

## 4 GoYoYo 的检索方法

### 4.1 检索词的输入

首先进入 GoYoYo 站点，使用中文国标码的用户可直接键入 http://www.goyoyo.com.hk/indexhkgb.html 进入下图状态，在图中上方的框内直接输入中文关键词，也可先选择类目再键入关键词。

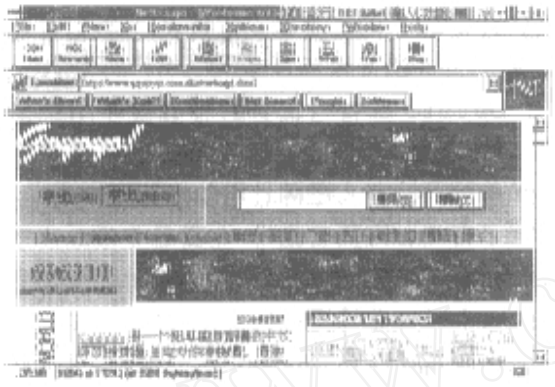

图 2 图中方框内输入检索词

词与词之间不必加空格，英文的个别单词也可以用。如果发现查找结果与你期望的不同，可以换个说法再试一次，也可以加个标点符号试试。须注意的是：

a. 不要使用范围太大的概念。如中国，这个词不能说明任何特别的主题，因为 GoYoYo 搜集主页绝大部分都与中国有关。

b. 注意概念的划分。尽量输入单一概念，避免使用组合的概念，因为 Goyoyo 对组合概念容易分得太细而不着边际。

c. 一般不要检索人名或个人的主页。如果这个人很有名，则另当别论。

### 4.2 检索限定

在检索窗口中输入逻辑函数，可以只列出一些特定的主页地址，系统会自动检查网址中是否包含，或不包含用户给定的字串。其主要的限定有以下几种：

1). hk 可以列出网址中带有“.hk”的主页。

2) GB 可以列出国标码的主页。

3) BIG5 可以列出大五码的主页。

4). edu AND GB 可以列出网址中带有“.edu”并且为国标码的主页。

5). edu OR .com 可以列出网址中带有“.edu”或者“.com”的主页。

6). edu NOT .cuhk 可以列出网址中带有“.edu”但没有“.cuhk”的主页。

7) NOT .cuhk.hk NOT .cuhk.edu.hk 可以列出网址中不带有“.cuhk.hk”和“.cuhk.edu.hk”的主页

8). sg OR .hk OR .tw 可以列出网址中带有“.sg”或者“.hk”或者“.tw”的主页。

在上述特定函数的限定中，可以使用 AND、OR、NOT 这几个布尔逻辑算符进行检索，但是注意以下两点：

1) AND、OR、NOT 都必须大写，而网址字串必须小写。有些网址中含有大写字母，但是在匹配时系统会自动按小写字母处理。

2) 逻辑运算是从左到右进行，没有优先级和括号。

### 4.3 检索结果的内容与格式

检索请求输入后，系统通常只需经过数秒钟的搜索就会马上给出检索结果，一般一条记录包含以下内容：

a. 你的检索请求；b. 有效的关键词，系统可识别的关键词；c. 找到的网页数目；d. 网页的标题、网页的中文代码、网页与你的请求的关系；e. 网页的 URL；f. 网页的起始一百多字符的内容；g. 网页的大小，分为 HTML 文件的大小，有用文字的字符数及中文的字符数；h. 网页的内容指数，此数越大表明此网页的内容越多并且与本站收录的网页内容越相近；i. 网页的最近更新时间；j. 网页的状态，如果有些项，说明已经有多次没有成功地访问到它了，因此它现在也可能就连不上；k. 相关网页。有时可以通过这里找到很多有用的网页。

## 5 GoYoYo 使用效果评价

笔者通过实践感到 GoYoYo 目前尚有某些较为明显的不足：

(1) GoYoYo 通常把词划分到最小单元，因而对一些组合概念无能为力。如：“防火墙”，就将它划分为“防”、“火”、“墙”三个词，这样检出的结果凡是带有这三个词的信息都出来了，拿到手里一看，全是一些与“防火墙”内容无关的东西；又比如“消防车”它划分为

图 3　检索结果显示

“消防”和“车”的两个概念, 结果把消防和各种汽车在一起的信息都检出来了, 而与“消防车”本身有着的信息则很少。

(2) 无法表示词与词之间的远近亲疏关系。GoYoYo 宣称它不像一般的检索系统那样仅仅进行简单的关键词匹配, 而是尽可能给你一个像样的结果, 并且还可以教导其智能机器系统一些知识, 告诉它那些东西是有联系的, 那些东西是无关的。事实上它还暂时无法做到这一点。笔者专门做过测验, 选取了“光纤”、“计算机”、“通信”和“网络”四个词进行组合, GoYoYo 将这四个词均作为有效词, 其中光纤和计算机 1999 条, 光纤计算机通信 1767 条, 光纤计算机通信网络 2000 条, 从检索结果看东西倒是不少, 但四个词之间的关系并不密切, 虽然词频增加了, 检索结果加大了, 但词与词之间的关系反而更加扑朔迷离, 内容一点也不明确, 像是一堆信息大杂烩, 有些还出现较大的误差。无法象布尔逻辑式关键词检索那样通过位置算符确定词与词的位置和关系, 如想检索与“光纤通信”相关的信息或与“网络计算机”相关的信息准确命中较为困难。

GoYoYo 建立的目标是想成为一个高度智能化的信息搜索工具, 检索语言也力求简化, 这个愿望当然是好的, 从实际使用效果来看, 目前对于须进一步细分的概念的确有效, 而且能扩大检索范围, 保证查全率。但对一些系统过于细分的概念词, 如前面提到的“防火墙”、“消防车”、“网络计算机”等, 其检索效果明显没有采用不加细分的关键词的命中率高。

GoYoYo 搜索引擎的出现, 无疑给 Internet 上中文信息的检索带来极大的益处, 尤其随着今后网上中文信息资源的日益丰富, 该检索系统更是大有可为, 但从用户的实际使用效果来看, 该搜索引擎确实还有待进一步完善。

## 参考文献


1　http://www.goyoyo.com.hk/

2　http://www.unilinx.com/


(上接第 9 页)

Wais(广域信息服务)

利用它能寻找 Internet 上分布的大量数据库

Usernet(网络用户论坛)

利用它能结识许多朋友, 并可获取同行或自己感兴趣的信息。

Veronica 和 Jughead

它们能维护许多 GoPher 菜单, 利用它可漫游 GoPher 空间, 然后使用户自动进入正确的 GoPher。

值得注意的是, Internet 资源虽无量, 且免费资源很多, 但是很多有价值的信息只提供一个样品, 其目的是引导你订购他们的信息资料。

## 参考文献


1　顾震宇. 如何在 Internet 上检索有用信息. 情报理论与实践, 1997, 2

2　胡小菁. 网络时代高校图书馆的职能与责任. 大学图书馆, 1996, 14(3)

3　谢新洲. 电子出版物知识讲座. 情报理论与实践, 1997, 2

4　毕德　邱枫等. 学用 Internet. 中国水利水电出版社, 1997, 1